平成２５年度 学内研究助成金 研究報告書

研　究　種　目	■奨励研究助成金	□研究成果刊行助成金
	□21 世紀研究開発奨励金 （共同研究助成金）	□21 世紀教育開発奨励金 （教育推進研究助成金）
研 究 課 題 名	クラスタ数自動推定アルゴリズムの開発 ～特に、情報量基準の観点から～	
研究者所属・氏名	研究代表者：理工学部　情報学科　講師　濵砂 幸裕 共同研究者：	

１．研究目的･内容

クラスタリングとは、大規模・複雑なデータの特徴を把握するデータ解析手法である。本申請課題では、クラスタリング分野における重要課題の１つである**最適クラスタ数決定問題の解決**を目的とし、**クラスタ数自動推定アルゴリズム**の開発を行う。特に、クラスタリング結果の評価に用いられる**妥当性基準**と確率モデルの評価に用いられる**情報量基準**の両者を対比し、それらの特性を明らかにすることで、より実用的なクラスタリング手法の開発を行う。

２．研究経過及び成果

①本申請課題遂行のための研究項目

本申請課題遂行のための研究項目は大きく次の３点にまとめることができる。

（１）情報量基準を用いたクラスタリング手法のモデル構築と評価
（２）クラスタ数自動推定アルゴリズムの開発
（３）実データを用いた数値実験による性能評価およびフィードバック

以降、上記の研究項目に従い、これまでの研究経過を報告する。

（１） 情報量基準を用いたクラスタリング手法のモデル構築と評価

本申請課題の目的であるクラスタ数自動推定アルゴリズムは、クラスタ分類と分類結果の評価を行うことで構成される。クラスタリング結果の定量的な評価を行う主な方法として、妥当性基準を用いる方法と情報量基準を用いる方法が知られている。両者の違いは妥当性基準が確率モデルを必要としないのに対して、情報量基準は分類結果に確率モデルを当てはめて評価を行う点である。本申請者は本研究課題着手より以前に、妥当性基準を用いたクラスタリング結果の評価に取り組んだ。また、情報量基準によるクラスタ数推定の対比を目的とし、情報量基準を用いたファジィ c-回帰モデルのクラスタ数推定を行った(第 57 回システム制御情報学会研究発表講演会にて発表、2013 年 5 月)。この中で、確率モデルによるクラスタリング手法である混合正規分布を用いた手法と複数のファジィ c-回帰モデルを比較したところ、確率モデルの有無に関わらず、クラスタ数推定の精度に大きな差異は見られなかった。情報量基準の算出に必要な確率モデルの推定には多くの計算時間が必要となるため、実用的には妥当性基準の利用が有効であることが考えられる。これらの結果より、情報量基準だけでなく妥当性基準による評価も想定し、複数のアルゴリズムを開発・比較することが本申請課題の達成に不可欠であると考えた。

（２）クラスタ数自動推定アルゴリズムの開発

（１）の検討をもとに、クラスタ数自動推定アルゴリズムとして、クラスタを逐次的に抽出する逐次抽出型アルゴリズム、k-means を再帰的に繰り返す X-means の改良および新規手法開発を進めた。はじめに、L1 正則化エントロピー型可能性クラスタリングを提案し、それを用いた逐次抽出型アルゴリズムを構築した（第 29 回ファジィシステムシンポジウムにて発表、2013 年 9 月)。本手法はノイズパラメータを用いる従来の逐次抽出型クラスタリング手法と等価なモデルを構成することが可能であるだけでなく、提案手法のパラメータを調節することにより従来手法よりも複雑な分類境界を構成することが可能である。

次に、k-means を再帰的に繰り返す X-means の改良を検討した。従来の X-means は k-means を再帰的に繰り返し、クラスタ分割の判定にベイズ情報量基準などを用いるため、確率モデルの推定が必要となる。そのため、計算時間の削減を目的とし、近似を用いた改良法が提案されている。ここで、（1）で得られた知見をもとに情報量基準の代わりに妥当性基準を用いた X-means の検討を行った。妥当性基準は確率モデルを必要としないため、クラスタ分類の結果を用いることで指標の値を算出することが可能である。そのため、情報量基準と比較すると計算時間がかなり少なくなり、実用に適した手法と言える。クラスタ分類の結果を定量的に評価する妥当性基準は代表的なものがいくつか知られているが、どの基準がどのような場合に有効であるかの明確な判断基準は存在しない。そこで、複数の妥当性基準を用いた X-means のアルゴリズムを構築した。現状では、構築したアルゴリズムを計算機上に実装し、様々な人工データや実データによる数値実験を行い、比較・検証を進めている段階である。

（3）実データを用いた数値実験による性能評価およびフィードバック

（2）で構築したクラスタ数自動推定アルゴリズムを用いて、 UCI Machine Learning Repository 上で公開されているベンチマークデータを用いて数値実験を行った。そこで得られた結果を評価・検討し、アルゴリズムの更なる高度化に取り組んでいる。具体的には、新たなアルゴリズムの提案、妥当性基準の検討ならびに関係データへの拡張などに取り組み、実世界のデータを効率的に解析するアルゴリズムの構築を進めている。今後は、モデル構築・アルゴリズム実装・数値実験による検証・フィードバックの PDCA サイクルを繰り返すことで、より実用的なクラスタ数自動推定アルゴリズムの開発を進める。

3．本研究と関連した今後の研究計画

②今後の研究計画

本研究課題の目的および得られた成果から、今後の研究計画として以下の 2 点を想定している。

（1） 逐次抽出型アルゴリズムの新規開発とプライバシー保護データマイニングへの応用

（2） 妥当性基準を用いた X-means の構築および関係データへの拡張

特に、Twitter などのソーシャルメディア上に蓄積された大規模データを扱い、本研究課題で提案したクラスタリング手法が有用なデータマイニングのツールとなることを示す。また、平成 27 年度中に査読付き学術論文雑誌へ掲載されるよう、研究成果をまとめ、本研究課題を総括する。

③本研究課題の自己評価

本研究課題の目的は**最適クラスタ数決定問題の解決**であり、目的達成のために複数のクラスタ数自動推定アルゴリズムの構築・検討を行った。本研究課題の最終的な自己評価は、

（1） 確率モデルを包含した新たな逐次抽出型アルゴリズムの構築

（2） 妥当性基準を用いた X-means の確立

以上の 2 点から行うこととなる。いずれか 1 つでも達成できれば、十分な成果と考えているが、本研究課題の成果を実問題に適用し、より波及させることを目標として設定している。

4．成果の発表等

発　表　機　関　名	種類（著書・雑誌・口頭）	発表年月日(予定を含む)
第 57 回システム制御情報学会研究発表講演会	口頭（査読なし）	2013 年 5 月 17 日
第 29 回ファジィシステムシンポジウム	口頭（査読なし）	2013 年 9 月 9 日