

A-07. クラスタ数推定に用いる最適な情報量基準の探求

萩原 涼介
(藤田 一寿)

1. はじめに

クラスタリングとはデータを教師なし学習により任意の数のクラスタに分ける手法である。k-meansを始めとする多くのクラスタリング手法では、予めクラスタ数がわかっているものとして、クラスタ数を指定しクラスタリングを行う。しかし、データに対し最適なクラスタ数を指定しなければ、最適なクラスタリング結果を得ることはできない。しかし、一般にクラスタ数が事前にわかっていない。その為、クラスタ数を推定することは重要な課題となっている。クラスタ数推定を行う際、よく用いられるのが情報量基準と呼ばれる指標である。情報量基準とは簡単に言えば確率分布とデータの分布の当てはまり具合を表すものである。その情報量基準は多くの研究者により様々なものが提案されている。しかし、どの情報量基準がどのようなデータに対し有効かは分かっていない。そこで本研究では、クラスタ数推定に用いる情報量基準として最適なものを数値実験を通し明らかにする。

2. 実験手法

本研究では X-means と呼ばれる手法を用い、クラスタ数推定およびクラスタリングを行った。X-means は情報量基準を用い、クラスタ数を決定する。AIC, cAIC, BIC と呼ばれる 3 つの情報量基準をそれぞれ用いクラスタ数推定およびクラスタリングを行い、その結果の比較を行った。精度の評価には、正規化相互情報量 (NMI) および Purity を用いた。それぞれの指標は 1 に近づくほど良いクラスタリング結果であると言える。

3. 実験結果

前期の数値実験では分散 $\sigma^2 = 1$ の 2 次元混合等方 Gauss 分布をデータセットとして用いた。このデータセットは 5 つの等方 Gauss 分布で構成される。そして各クラスタは 500 個のデータ点を持つ。このデータセットに対し、X-means によるクラスタリングを行った。

表 1 は推定したクラスタ数とクラスタリングの精

度を表す。それぞれの数値は 100 回実行した結果を平均したものである。この結果、混合等方 Gauss 分布では BIC と cAIC が適していることがわかった。

また、後期は実データとして手書き数字データセット (MNIST) のクラスタ数推定を行った。比較のため、確率ベースでないクラスタ数推定を行う手法として Mean shift によるクラスタ数推定も行った。

結果を表 2 に示す。その結果、X-means を採用した場合も Mean shift を採用した場合も適切なクラスタ数推定の結果を得ることができなかった。

表 1 2 次元空間におけるクラスタリング結果

| 分割停止基準 | クラスタ数 | NMI | Purity |
|--------|-------|---------|---------|
| BIC | 4.58 | 0.88281 | 0.84459 |
| cAIC | 4.55 | 0.89993 | 0.85329 |
| AIC | 4.69 | 0.88147 | 0.83642 |
| 対数尤度関数 | 5.32 | 0.91572 | 0.85700 |

表 2 手書き数字データのクラスタリング結果

| クラスタリング手法 | クラスタ数 |
|------------------|-------|
| X-means (AIC) | 32 |
| X-means (cAIC) | 32 |
| X-means (BIC) | 32 |
| X-means (対数尤度関数) | 32 |
| Mean shift | 1 |

4. おわりに

混合等方 Gauss 分布から生成される人工データのクラスタリングには BIC が最も適していることがわかった。実データのクラスタ数推定を行う際には、そのデータにあったモデルやクラスタ数推定の手法を採用する必要がある。実データのクラスタ数を適切に推定するための手法を検討していく必要がある。

【研究目標】

クラスタ数推定に用いる最適な情報量基準を探求する

【課題項目】

○: 数値実験, △: 実データのクラスタリング