非階層的クラスタリングにおけるクラスター数の決定法

Determining the number of clusters in non-hierarchical clustering

数学専攻　柴田　祐大

SHIBATA, Yoshihiro

1　はじめに

クラスター分析とは, 複数の変数で特徴づけられた多数の個体を, 個体間の類似性の基準に基づいていくつかのクラスに分類するための手法であり, クラスタリングとも呼ばれている. マーケティングの分野では顧客の購買行動の特性などに基づいて, グループ化することにより特徴を捉えやすくしたり, 特徴が似ているものは同じような購買行動に繋がるものとして施策やターゲティングに利用されている. 近年では, コンピューター処理能力の格段の向上により, 大量のデータが取得可能となり, クラスター分析は様々な分野で有用な解析方法となってきた. さらに, 分析技術の向上により, 文章や画像のデータも処理することができるようになり, その処理過程において, 事前のデータの前処理や分析した多くのデータをグループ分けしたり, 可視化する手法が注目されてきた.

非階層的クラスタリングの代表的な手法として, MacQueen (1967) による k-means 法や Kohonen (1995) による自己組織化マップがあるが, これらの手法は事前にクラスター数を与える必要がある. 実際には, クラスター数は未知であるため, クラスター数を事前にどのような方法によって決めるかという問題が生じる. この問題に対して, Pelleg and Moore (2000) は, 混合分布モデルとベイズ型モデル評価基準 BIC (Schwarz, 1978) に基づくクラスター数の自動決定法である X-means 法を提案した. 石岡 (2000) は, Pelleg and Moore による X-means 法の計算アルゴリズムを改良し, 大規模データに対応できるアルゴリズムを提唱した. X-means 法は構成したクラスターの分割を更新するか否かを BIC の値の大小関係に基づいて行なっている. しかし, この方法では, BIC の値に僅かしか差がない場合に, 分割の判断基準として適切であるかという問題が生じる. この問題に対して, 本論文では, X-means 法の計算アルゴリズムの中に Linhart (1988) による情報量規準 AIC (Akaike, 1973) の差の検定を組み込んだクラスター数の自動決定法を提案した.

2　X-means 法

本節では, k-means 法で適切なクラスター数を決定する方法として Pelleg and Moore (2000) による X-means 法という方法について述べる. X-means 法は, ベイズ型モデル評価基準 BIC を用いることで, 十分に小さいクラスターから始めて, 各クラスターにおいて, 分割が適切であると判断されるまで, 2 分割を繰り返していく. Pelleg and Moore (2000) のアルゴリズムについて述べる.

STEP (0) :　分類対象とする n 個の p 次元データを $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ とする.

STEP (1) :　十分小さなクラスター数の初期値 k_0 (特に指定がなければ 2) を定める.

STEP (2) :　$k = k_0$ として, k-means 法を適用する. 分割後のクラスターを $C_1, C_2, \ldots, C_{k_0}$ とする.

STEP (3) :　この時点での総クラスター数を g とする. クラスター C_i に対して $k = 2$ として, k-means 法を適用する. 分割した後のクラスターを C_i^1, C_i^2 とする.

STEP (4) : 分割する前の場合について考える. すべてのデータに対しての p 変量混合正規分布を

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{j=1}^{g} \pi_j \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\right\} \tag{1}$$

と仮定する. ただし, $\sum_{i=1}^{g}\pi_i = 1$ とする. 分散共分散行列は, $\Sigma = \mathrm{diag}(\sigma^2)$ と仮定する. 分割する前の BIC を以下のように計算する.

$$\mathrm{BIC} = -2\log L(\boldsymbol{x};\hat{\boldsymbol{\theta}}) + q\log n$$

ここで, $\hat{\boldsymbol{\theta}} = [\hat{\pi}, \hat{\boldsymbol{\mu}}, \hat{\Sigma}]$ は, p 変量混合正規分布の推定量であり, q はパラメータ空間の次元数で, $q = g-1+gp+1 = g(p+1)$ である. $\boldsymbol{x}_i$ はクラスター C_i に含まれる p 次元データとする. L は尤度関数で $L(\cdot) = \prod_{i=1}^{n} f(\cdot)$ である.

STEP (5) : C_i^1, C_i^2 と 2 分割に分けた場合について考える. すべてのデータに対しての p 変量混合正規分布を

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{j=1}^{g+1} \pi_j \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\right\} \tag{2}$$

と仮定する. ただし, $\sum_{i=1}^{g+1}\pi_i = 1$ とする. 分散共分散行列は, $\Sigma = \mathrm{diag}(\sigma^2)$ と仮定する. この 2 分割モデルにおける BIC を以下のように計算する.

$$\mathrm{BIC}' = -2\log L(\boldsymbol{x};\hat{\boldsymbol{\theta}}) + q'\log n \tag{3}$$

パラメータ数は, $q' = g+(g+1)p+1 = (g+1)(p+1)$ となる.L は尤度関数で $L(\cdot) = \prod_{i=1}^{n} f(\cdot)$ である.

STEP (6) : BIC > BIC' ならば, 2 分割モデルが適切であると判断して, 2 分割を継続すべく $C_i \leftarrow C_i^1$ とする. C_i^2 については, p 次元データ, クラスターの重心, 対数尤度とモデル評価基準を保持して, STEP (3) へ戻る.

STEP (7) : BIC < BIC' ならば, 2 分割しないモデルの方がより適切であると判断され, C_i についての 2 分割を停止する.

STEP (8) : C_i における 2 分割が全て終了. STEP (4) $\sim$ (7) で作成された 2 分割クラスターが C_i 内で一意になるようにデータに属するクラスター番号を振り直す.

STEP (9) : $i = 1, 2, \ldots, k_0$ に STEP (4) $\sim$ (8) を繰り返す.

STEP (10) : 最初に k_0 個に分割したクラスター全てにおいて 2 分割が全て終了. 全てのデータに対して一意になるようにデータに属するクラスター番号を振り直す.

これが, Pelleg and Moore (2000) による X-means 法のアルゴリズムである. クラスター数を決定するためにこのアルゴリズムでは, 分割の基準としているモデル評価基準の値の大小を単純に比べているだけなので, クラスター数の推定に誤差が生じる可能性がある. 本論文では, モデル評価基準の値の差を単に比較するだけではなく, 差の有意性の検定をアルゴリズムに組み込んだ方法を提案した. このため, Linhart (1988) による情報量規準 AIC の差の検定を用いた. また, k-means 法でデータの共分散を考えていないため, 共分散があるデータに対してうまくクラスタリングができていない. そのため, 単純に距離を比較するだけでなく, 共分散を考慮したマハラノビス距離を用いる. 次節で提案したアルゴリズムを述べる.

3 X-means 法の改良

本節では, X-means 法を改良してより一般化した形で表し, 複雑なデータに対してもより適切にクラスター数を推定可能なアルゴリズムを提案する. X-means 法からの改良点としては, k-means 法でのクラスタリングの時に, クラスターの核から各特徴量の距離の測り方をマハラノビス距離を用い, クラスターの共分散構造を捉えられるようにする. また, モデル選択においても, モデル評価基準の値を比較するだけではなく, Linhart (1988) で提唱されたモデル評価基準の差の検定を行い, モデル評価基準のサンプリングエラーを評価した. それにより, 一般的なデータについて汎用的なアルゴリズムを提案する. ここでは, Pelleg and Moore (2000) のアルゴリズムを元に提案手法の変更点について述べる.

STEP (2), STEP (3) : この STEP で用いる k-means 法の距離計算の式に関して, クラスターの核と各データ間の距離に対して, 以下のマハラノビス距離を用いる.

$$D = (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{4}$$

クラスターの標本平均ベクトルを $\boldsymbol{\mu}$, 標本分散共分散行列を Σ として, 分類対象とするデータを $\boldsymbol{x}$ とする.

STEP(4) : 分割する前について考える. 分割する前のクラスター数を g とする. すべてのデータに対しての p 変量混合正規分布を

$$f_1(\boldsymbol{x}; \Theta_1) = \sum_{j=1}^{g} \pi_j \frac{1}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_j) \Sigma_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \right\} \tag{5}$$

と仮定する. ここで, $\sum_{j=1}^{g} \pi_j = 1$ で, $\hat{\Theta}_1 = \hat{\boldsymbol{\theta}}_{\boldsymbol{j}} = [\hat{\pi}_j, \hat{\boldsymbol{\mu}}_j, \hat{\Sigma}_j]$ は, p 変量混合正規分布の EM アルゴリズムによる推定量とする. q はパラメータ空間の次元数で, $q = \frac{g}{2}(p+1)(p+2) - 1$ である. $\boldsymbol{x}_i$ はクラスター C_i に含まれる p 次元データとする. L_1 は尤度関数で $L_1(\cdot) = \prod_{i=1}^{n} f_1(\cdot)$ である. 確率密度関数, 尤度関数, パラメータ数は, STEP (6) で使用するまで保持しておく.

STEP(5) : C_i^1, C_i^2 と 2 分割した場合について考える. すべてのデータに対しての p 変量混合正規分布を

$$f_2(\boldsymbol{x}; \Theta_2) = \sum_{j=1}^{g+1} \pi_j \frac{1}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_j) {\Sigma_j}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \right\} \tag{6}$$

と仮定する. パラメータ数は, $q' = \frac{(g+1)}{2}(p+1)(p+2) - 1$ となる. L は尤度関数で $L_2(\cdot) = \prod_{i=1}^{n} f_2(\cdot)$ である. 確率密度関数, 尤度関数, パラメータ数は, STEP (6) で使用するまで保持しておく.

STEP(6) : このステップでは, モデル評価基準の検定を行う (Linhart, 1988; 下平, 1999) モデル評価基準のサンプリングエラーを評価するためには, 二つのモデルの最大対数尤度の差の分散を

$$\hat{\sigma}_{1,2}^2 = \sum_{t=1}^{n} (\log f_1(\boldsymbol{x}_t | \hat{\boldsymbol{\theta}}_1) - \log f_2(\boldsymbol{x}_t | \hat{\boldsymbol{\theta}}_2))^2 - \frac{1}{n} (\log L_1(\boldsymbol{x}; \hat{\Theta}_1) - \log L_2(\boldsymbol{x}; \hat{\Theta}_2))^2 \tag{7}$$

と推定する. また, 分割前の情報量規準 AIC を以下のように計算する.

$$\mathrm{AIC}_1 = -2 \log L_1(\boldsymbol{x}; \hat{\Theta}_1) + 2q$$

2 分割モデルにおける AIC を以下のように計算する.

$$\mathrm{AIC}_2 = -2 \log L_2(\boldsymbol{x}; \hat{\Theta}_2) + 2q'$$

漸近的に正規分布 $N(0,1)$ に従う統計量

$$T_{1,2} = \frac{\mathrm{AIC}_1 - \mathrm{AIC}_2}{2\hat{\sigma}_{1,2}} \tag{8}$$

を使って, 情報量規準の差の有意性を検定する. 標準正規分布 $N(0,1)$ の分布関数を $\Phi(x)$ とした時, $T_{1,2} > \Phi^{-1}(1-P*)$ となる時, 近似的に有意水準 $P*$ で 2 分割モデルの方が適切であると判断して, 2 分割を継続すべく $C_i \leftarrow C_i^1$ とする. C_i^2 については, p 次元データ, クラスターの重心, 対数尤度と情報量規準を保持して, STEP (3) へ戻る.

STEP(7) : $T_{1,2} < \Phi^{-1}(1-P*)$ ならば, 2 分割しない方が適切であると判断され, C_i についての 2 分割を停止する.

提案手法では, クラスター分割アルゴリズムのプロセスの中で, 情報量規準 AIC の大きさを単純に比較するのではなく, 情報量規準の差の検定を組み込んだ. 2 分割する前の AIC が 2 分割した後の AIC 以下であると帰無仮説を置いて検定を行なっている. もしも, $\mathrm{AIC}_1 > \mathrm{AIC}_2$ であったとしても, $T_{1,2} > \Phi^{-1}(1-P*)$ ならば, 帰無仮説を棄却することが出来ないため, 2 分割する前の AIC と 2 分割した後の AIC に必ずしも有意な差があるとは言えないと判断されるため, 情報量規準のサンプルエラーを考慮しながら, クラスター数の推定ができる. また, STEP (2), (3) でマハラノビス距離を用いることにより, データの散らばりに対応した適切なクラスタリングをすることができる.

4 おわりに

本論文と Pelleg and Moore (2000) の提案した X-means 法との大きな違いは, 分割の可否を決定するモデル評価基準の値の差の大小関係を単に比較するのではなく有意性の検定に置き換えた点にある.

今後, 提案したアルゴリズムでどの程度の精度でクラスター数を推定できるかを数値実験によって検証する. また, 研究課題として, 今回提案した X-means 法の改良方法では, 初期値におけるマハラノビス距離の測り方, 有意水準 $P*$ の決め方, BIC に基づく差の検定法を求め, AIC と BIC のどちらを使用した方が精度が高いか数値的に研究することなどが挙げられる.

参考文献

[1] 石岡恒憲 (2000). クラスター数を自動決定する k-means アルゴリズムの拡張について. 応用統計学. **29** 141-149.

[2] 小西貞則・北川源四郎 (2004). 情報量規準. 朝倉書店.

[3] 小西貞則 (2010). 多変量解析入門 — 線形から非線形へ. 岩波書店.

[4] Linhart, H (1988). A test whether two AIC's differ significantly. *South African Statist. J.* **22**, 153-161.

[5] Pelleg, D. and Moore, A (2000). X-means: Extending K-means with efficient estimation of the number of clusters. ICML-2000.

[6] 下平英寿 (1999). モデル選択理論の新展開. 統計数理. **47**, 3-27.