D-020

x-means 法における分割停止規準の改良

嘉村準弥 † 小柳滋 ††
† 立命館大学大学院理工学研究科 †† 立命館大学情報理工学部

1 はじめに

現在，クラスタリングの大半の手法は分析者がクラスタ数を決定しなくてはならない．その問題を解決するため，自動的に適切なクラスタ数を決定する手法として x-means 法が考案されている．x-means 法は k-means 法を用いて分割を繰り替えし実行する手法であり，情報量規準を用いることによって分割を停止させることによりクラスタを決定する．本研究では x-means 法の情報量規準に焦点を当て，様々な規準を用いることにより，精度の高いクラスタリングを目指す．

2 x-means 法の概要

2.1 x-means 法

x-means 法は Pelleg and Moore[1] によって提案された非階層型クラスタリングの一種である．x-means 法は k-means 法を拡張したものであり，クラスタ数 k を自動的に定めるという点が特徴である．x-means 法の基本的な動作は，2 分割する k-means アルゴリズムを情報量規準に沿って妥当な限り繰り返すという単純なものである．本稿においては石岡氏 [2] により改良された x-means 法を使用した．これは逐次分割されるクラスタごとに分散の違いを考慮した点と，BIC の計算に用いる対数尤度の計算に近似計算を行って計算速度を向上させている点が改良されている．

2.2 アルゴリズム

n 個 p 次元のデータを扱う場合を想定する．

1. 与えられたデータに対して小さい分割クラスタ数 k_0 を与える
2. $k = k_0$ として k-means クラスタリングを行う．分割後のクラスタを $C_1, C_2, ..., C_{k0}$ とする
3. $i = 1, 2, ..., k_0$ としてスタックに積んでいき，以下の手順を行う．
4. スタックより取り出したクラスタ C_i に含まれるデータに p 変量正規分布

$$f(\theta_i; d) = (2\pi)^{p/2}|V_i|^{1/2} exp\left[-1/2(d-\mu_i)^t V_i^{-1}(d-\mu_i)\right] \quad (1)$$

を仮定する．この分布についての BIC を以下のように求める．

$$BIC = -2\log L(\hat{\theta}_i; x_i \in C_i) + q\log n_i \quad (2)$$

ここで $\hat{\theta}_i = [\hat{\mu}_i, \hat{V}_i]$ は p 変量正規分布の最尤推定値とする．ここでは正規分布を仮定しているので $\hat{\mu}_i$ は相加平均，$\hat{V}_i$ は相乗平均と等しくなる．μ_i は p 次の平均値ベクトル，V_i は $p \cdot p$ の分散共分散行列である．q は独立なパラメータ空間の次元数，つまり p 次元の平均値ベクトルと $p \cdot p$ の分散共分散行列の組み合わせ数を表す．よって $q = p + p \cdot {}_pC_2 = (p+3)/2$ である．x_i は C_i に含まれている p 次元データの集合，n_i は C_i に含まれているデータ数とする．L は尤度関数 $L(\cdot) = \Pi f(\cdot)$ である．

5. C_i に対して k-means 法で 2 分割を行い，C_i^1，C_i^2 とする．
6. C_i^1, C_i^2 に対して，それぞれパラメータ θ_i^1,θ_i^2 をもつ p 変量正規分布を仮定し，2 分割モデルにおいてデータが従う確率密度とする．あるデータがどちらのパラメータを用いるかは属しているクラスタによって決まる．

$$x_i = \begin{cases} \alpha_i[f(\theta_i^1); x], & x_i \text{が } C_i^1 \text{に含まれるとき} \\ \alpha_i[f(\theta_i^2); x], & x_i \text{が } C_i^2 \text{に含まれるとき} \end{cases} \quad (3)$$

α_i は上式を確率密度とするための基準化定数で

$$\alpha_i = \begin{cases} 1/\int[f(\theta_i^1; x_i)]dx, & x_i \text{が } C_i^1 \text{に含まれるとき} \\ 1/\int[f(\theta_i^2; x_i)]dx, & x_i \text{が } C_i^2 \text{に含まれるとき} \end{cases} \quad (4)$$

となる．厳密に求めると p 次積分が必要となるため，多くの計算量が必要となるので近似値として

$$\alpha = 0.5/K(\beta_i) \quad (5)$$

を用いる．$K(\cdot)$ は標準正規分布の下側確率であり，β_i は $f(\theta_i^1; x_i)$ と $f(\theta_i^2; x_i)$ の分離の程度を示す指標で

$$\beta_i = \sqrt{\frac{||\mu_1 - \mu_2||^2}{|V_1| + |V_2|}} \quad (6)$$

Improvement of Division Criteria for x-means algorithm
†† Junya KAMURA
† Shigeru OYANAGI
College of Information Science and Engineering, Ritsumeikan University (†)

で示される．
これらを用いて 2 分割モデルにおける BIC を計算する．

$$BIC' = -2\log L(\hat{\theta}_i'; x_i \in C_i) + q' \log n_i \quad (7)$$

$\hat{\theta}_i = [\hat{\theta}_i^1, \hat{\theta}_i^2]$ は 2 つの p 変量正規分布の最尤推定値である．各々の p に対して分散と平均のパラメータが存在するので，パラメータ空間の次元数は $q' = 2q = p(P+3)$ となる．

7. BIC>BIC' ならば，2 分割したクラスタのほうが好ましいモデルと判断できるので結果クラスタフラグを偽に設定し，2 分割されたクラスタの保持データ，新たに選択されたセントロイド，それぞれの BIC をスタックに積む．

   BIC<BIC ' ならば，分割前クラスタのほうが好ましいモデルと判断出きるので，結果クラスタフラグを真に設定し，2 分割されたクラスタを破棄する．

   スタックが空で無かったら手順 4 へ，空なら次ステップへ進む．

8. 全てのクラスタの分割が終わり，x-means 法が終了する．本来ならクラスタ番号を一意に振り直したほうがよいが今回は分割の様子を観測するためにあえて行わない．結果クラスタフラグが新であるクラスタが x-means 法によって導き出されたクラスタ集合である．

3 提案手法

x-means 法における問題点として，分割停止の条件として情報量規準に BIC を用いている点であると考えている．従来の x-means 法に用いられている BIC は与えられるデータの性質によって好ましい値を示さない場合がある．これは BIC が p 変量正規分布をモデルとしているためノイズを多く含むデータに対してはクラスタ数を少なくとらえてしまう恐れがあるためである．この問題は他の情報量規準である AIC においても同様である．また，これらの規準は漸近理論で導出されており，標本数無限を仮定しているため，標本数が少数では偏りが生じてしまう．

本研究では分割停止規準の変更に伴うクラスタリング精度の向上を図る．分割停止規準に用いる指標として情報量基準ではなく Cluster validity Index（クラスタ妥当性指標）を使うことにより，クラスタリング結果の評価を行いながら好ましいクラスタ数を定める．以前のクラスタリング結果より評価が良好であれば分割を継続し，そうでなければ分割を停止させるアルゴリズムに変更する．

用いる妥当性指標の一つとして Davis-Bouldin Index（DBIndex)[3] がある．この指標は各クラスタがまとまっており，クラスタ間距離が離れているほどよいクラスタであるという前提に値を求める．DBIndex は以下のようにして求められる．Ω_i は i 番のクラスタ要素の集合，μ_i は平均を表す．

i 番のクラスタについてのまとまり s_i を求める式として分散を使用する．

$$s_i = \frac{1}{card(\Omega_i)} \sum_{x \in \Omega_i} ||x - \mu_i||^2 \quad (8)$$

クラスタ間距離 d_{ij} は重心のユークリッド距離によって求める．

$$d_{ij} = ||\mu_i - \mu_j||^2 \quad (9)$$

上記の式から i 番のクラスタと他のクラスタを比較した評価値 r_i を得る．比較するクラスタは最も大きい値を示すものとする．

$$r_i = max_{j,\, j \neq i} \frac{s_i + s_j}{d_{ij}} \quad (10)$$

全てのクラスタよりで得られた r_i 値の平均が DBIndex である r となる．r が小さいほどそのクラスタリングはよい結果であるといえる．

$$r = \frac{1}{c} \sum_{i=1}^{c} r_i \quad (11)$$

4 おわりに

現在，分割を行った後，クラスタリング結果の妥当性指標を求めるプログラムを組んでいるが，期待している値を示さないので改善している段階である．また，提案しているクラスタ妥当性指標は DBIndex のみであるが今後は他の指標である silhouette index などを用いた場合どのようになるか検証する．またこの 手法を適用し，検証するにあたり様々なデータセットにおいて有効であるか考察することが重要である．

参考文献

[1] Dan Pelleg, Andrew Moore (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters

[2] 石岡 恒憲, 『クラスタ数を自動決定する k-means アルゴリズムの拡張について』

[3] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan. Data Mining A Knowledge Discovery Approach. (pp280-284)