Group LassoによるFuzzy c-meansクラスタリングの変数選択

大阪大学大学院　宇野 光平

Fuzzy c-meansクラスタリングはk-meansクラスタリングと同様に非階層的クラスタリングであるが, ノイズ変数が増えるほどクラスター構造を見つけることが難しくなるため, 重要な変数のみを選ぶことはクラスタリングの精度向上のために必要である. そこで本研究では, Fuzzy c-meansクラスタリングの目的関数にセントロイドについてのGroup Lassoペナルティを加える. ペナルティによって, ある変数における各クラスターの代表点を示す値が全て0になる場合が生じる. セントロイドの値が全て0になった変数をクラスタリングには寄与しないノイズ変数とみなすことで, 変数選択を行う.

1. 導入

データ行列$\mathbf{X} \in \mathbb{R}^{N\times P}$が与えられ各個体$n = 1, ..., N$をグループ$c = 1, ..., C$に分類したいとき, Fuzzy c-meansクラスタリングの目的関数は,

$$f(\mathbf{X}, \mathbf{B}) = \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}^{\alpha}\|\boldsymbol{x}_n - \boldsymbol{b}_c\|_2^2 = \sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}^{\alpha}\sum_{p=1}^{P}(x_{np} - b_{cp})^2 \tag{1}$$

である. ただし, αはファジィさを調整するパラメータであり, $\boldsymbol{x}_n \in \mathbb{R}^P$は個体$n$のデータベクトルを表し, $\boldsymbol{b}_c \in \mathbb{R}^P$はクラスター$c$のセントロイドベクトルを表す.

しかし(1)式は, 各変数についての目的関数ではない. ある条件で変数pのセントロイドベクトル$\boldsymbol{c}_p$が, 0ベクトルになるようなペナルティ$P(\boldsymbol{c}_p)$を加えたいため, 目的関数を書き換える必要がある. そこで, Yamashita & Mayekawa (2015)で明示されている行列形式の目的関数を用いる.

2. モデル

提案手法の目的関数は以下である.

$$f_R(\mathbf{X}, \mathbf{B}) = \sum_{p=1}^{P}\|\mathbf{D}_{\mathbf{U}}(\mathbf{1}_C \otimes \mathbf{I}_N)\boldsymbol{x}_p - \mathbf{D}_{\mathbf{U}}(\mathbf{I}_C \otimes \mathbf{1}_N)\boldsymbol{b}_p\|_2^2 + \sum_{p=1}^{P}\lambda\|\boldsymbol{b}_p\|_2 \tag{2}$$

ただし$\mathbf{D}_{\mathbf{U}} = \mathrm{diag}(\mathrm{vec}(\mathbf{U}^{(\alpha/2)}))$であり, λはチューニングパラメータとする.

参考文献

Yamashita, N., & Mayekawa, S-I. (2015). A new biplot procedure with joint classification of objects and variables by fuzzy c-means clustering. *Advances in Data Analysis and Classification*, **9**, 243-266.