A-006

Mallows の C_p 規準による回帰式の変数選択：混合整数二次計画法を用いた解法
Subset Selection by Mallows' C_p: A Mixed Integer Programming Approach

高野 祐一 *1　　宮代 隆平 *2
Yuichi TAKANO　Ryuhei MIYASHIRO

1　はじめに

本発表では，線形回帰モデルの説明変数を選択する問題を扱う．回帰分析では，説明変数を増やせば，推定に使用するデータへのモデルの当てはまりは良くなる (少なくとも悪くはならない)．しかし，過度に複雑なモデルを作成すると，過剰適合によって事後的な予測性能が悪化してしまう．ゆえに，予測モデルとしての性能を向上させるためには，説明変数を適切に選択してシンプルかつ説明力のあるモデルを作成することが重要となる．

変数選択に利用される適合度指標として，赤池情報量基準 (AIC) やベイズ情報量基準 (BIC) などの情報量基準や自由度調整済決定係数 $\bar{R}^2$ がある．実用規模のデータセットに対して，これらの適合度指標を目的関数として最適な説明変数集合を求めることは非常に難しいが，Miyashiro and Takano [1] は混合整数二次錐計画法による厳密解法を提案した．論文 [1] では，ステップワイズ法との比較を通して提案手法の有効性を検証しているが，厳密解法であるがゆえに計算に時間を要するという欠点があった．

そこで，本発表では適合度指標として Mallows の C_p 規準 [2] を利用し，問題を混合整数二次計画問題に帰着する解法を提案する．

2　線形回帰モデル

本稿では，以下の線形回帰モデルを扱う:

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_px_p + \varepsilon, \qquad (1)$$

ただし，y は予測すべき被説明変数，x_j $(j = 1, 2, \ldots, p)$ は予測に用いる説明変数，a_j $(j = 0, 1, \ldots, p)$ は推定すべき切片と偏回帰係数，ε は予測残差である．上記のモデルは選択候補となるすべての説明変数を使用しており，フルモデルと呼ばれる．

ここで，回帰モデルの推定のために n 種類のサンプル $(y_i; x_{i1}, x_{i2}, \ldots, x_{ip})$ $(i = 1, 2, \ldots, n)$ を用意して，$\boldsymbol{y} := (y_1\ y_2\ \cdots\ y_n)^\top$, $\boldsymbol{a} := (a_0\ a_1\ \cdots\ a_p)^\top$, $\boldsymbol{\varepsilon} := (\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n)^\top$,

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

とすると，フルモデル (1) は以下のように書き直せる:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\varepsilon}.$$

そして，残差二乗和 $\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}$ を最小化する最小二乗推定量 $\hat{\boldsymbol{a}}$ は以下で表される:

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

3　Mallows の C_p 規準

Mallows の C_p 規準 [2] は，主に線形回帰モデルに使われる適合度指標である．また，線形回帰モデルに対しては AIC と C_p の最小化はほぼ等価であることも知られている [3]．予測残差 ε_i $(i = 1, 2, \ldots, n)$ が平均 0，分散 σ^2 の独立同分布に従うと仮定する．ここで，選択する説明変数の添字集合を S $(\subseteq \{1, 2, \ldots, p\})$ で表すことにし，偏回帰係数の値を 0 にすることは対応する説明変数を回帰式から削除することと等しいことに注意すると，説明変数集合として S を選択した回帰モデルの C_p は以下のように定義される:

$$\min_{\boldsymbol{a}} \left\{ \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})}{\hat{\sigma}^2} \;\middle|\; a_j = 0\ (j \notin S) \right\} + 2(|S| + 1) - n,$$

ただし，$\hat{\sigma}^2$ は残差分散 σ^2 の推定量であり，フルモデル (1) の残差分散の不偏推定量

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{a}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{a}})}{n - p - 1} \qquad (2)$$

が使われることが多い [2, 3].

定数部分を除去すると，C_p は残差二乗和 (モデルの当てはまりの良さ) と説明変数の数 (モデルの複雑さ) のバランスを考慮した指標となっていることが分かる．また，推定量 (2) の下でフルモデルの C_p は $p + 1$ となるため，C_p 規準の変数選択によって C_p の値は $p + 1$ 以下となることも分かる．

4　定式化

Mallows の C_p 規準による変数選択問題を混合整数二次計画問題として定式化する．まず，j 番目の説明変数を選択する/しないを表す 0-1 決定変数

$$z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p) \qquad (3)$$

を導入する．そして，$z_j = 0$ の場合には以下の制約条件によって説明変数を回帰式から削除する:

$$z_j = 0 \ \Rightarrow\ a_j = 0 \quad (j = 1, 2, \ldots, p). \qquad (4)$$

*1 専修大学，Senshu University
*2 東京農工大学，Tokyo University of Agriculture and Technology

このような論理条件は分枝限定法の分枝操作の中で扱うことができ，例えば整数計画ソルバー CPLEX の `indicator` 関数を使えばこの機能が利用できる．またよく知られている big-M 法でも，このような論理条件は表現できる．

選択された説明変数の数が $\sum_{j=1}^{p} z_j$ と等しいことに注意すると，C_p 規準による変数選択問題は以下のように定式化できる:

$$\begin{aligned} \underset{\boldsymbol{a},\boldsymbol{z}}{\text{最小化}} \quad & \frac{\sum_{i=1}^{n}\left(y_i-\left(a_0+\sum_{j=1}^{p}a_j x_{ij}\right)\right)^2}{\hat{\sigma}^2} \\ & +2\left(\sum_{j=1}^{p} z_j+1\right)-n \\ \text{制約条件} \quad & (3),(4). \end{aligned}$$

この定式化は線形制約・凸二次目的関数の混合整数二次計画問題であり，論文 [1] で提案された定式化 (混合整数二次錐計画問題) と比較して非常に高速に解くことができる．

5 計算実験

UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml`) からダウンロードしたデータセットを使用して計算実験を行なった．なお，混合整数凸二次計画問題の求解には CPLEX を使用し，データ解析ソフト R のステップワイズ法と性能を比較した．結果を要約すると，

- 選択候補変数が 30 個以下であれば，提案手法は数秒程度で最適解を得ることができた．
- 選択候補変数が 100 個以下であれば，提案手法は 1000 秒以内にステップワイズ法と同等以上の暫定解を得ることができた．
- 選択候補変数が 400 個程度の場合は，提案手法は 10000 秒以内にステップワイズ法と同等以上の暫定解を得ることができた．
- サンプル数が多い (10000 以上の) 場合は，提案手法はステップワイズ法と同等以上の暫定解をステップワイズ法よりも短時間で得ることができた．最も顕著な問題例では，ステップワイズ法が 30000 秒以上かけて求めた解を上回る解を，提案手法では 1800 秒程度で見つけることができた．

計算結果の詳細については当日の発表と Miyashiro and Takano [4] を参照されたい．

6 おわりに

本発表では，Mallows の C_p 規準による回帰式の変数選択のための解法を提案した．整数計画ソルバーの性能は急速に進歩しており，本発表の計算実験を通して，最新鋭の整数計画ソルバーが統計的モデル選択のための実用的なツールとなりえることを実証することができたと考える．

参考文献

[1] R. Miyashiro and Y. Takano, “Mixed Integer Second-Order Cone Programming Formulations for Variable Selection,” Technical Report No. 2013-7, Department of Industrial Engineering and Management, Tokyo Institute of Technology (2013).

[2] C.L. Mallows, “Some Comments on C_p,” *Technometrics*, Vol. 15, No. 4, pp. 661–675 (1973).

[3] A. Miller, *Subset Selection in Regression, 2nd Edition* (Chapman and Hall/CRC, 2002).

[4] R. Miyashiro and Y. Takano, “Subset Selection by Mallows’ C_p: A Mixed Integer Programming Approach,” Technical Report No. 2014-1, Department of Industrial Engineering and Management, Tokyo Institute of Technology (2014).