4A-04

確率的離散一次法による一般化線形モデルの特徴選択

工藤晃太† 高野祐一‡

†筑波大学システム情報工学研究群 ‡筑波大学システム情報系

1 はじめに

特徴選択とは，予測モデルを構築する際に，利用可能な特徴量の中から有効な特徴量を選択することである．特徴選択によって，データ収集や保管のコストの削減，パラメータ推定の高速化，因果関係の理解の促進，過剰適合の軽減による予測性能の向上などの利点がある．

大規模な特徴選択問題を解くため，多くの先行研究では発見的解法が利用されている．近年 Bertsimas et al. [1] は，線形回帰モデルの特徴選択問題を効率的に解くアルゴリズムとして，離散一次法を提案した．この手法は，凸最適化で用いられる勾配降下法を離散最適化に応用したものであり，局所的最適解への収束性が証明されている．離散一次法を改良した確率的離散一次法 [2] は，探索点列に確率変動を加えることで局所的最適解を脱出し，さらに目的関数に L2 正則化項を導入することで高い予測性能を実現している．

しかし，離散一次法は線形回帰モデルを対象としており，一般化線形モデルではアルゴリズムの実行に必要となる目的関数の勾配に対するリプシッツ定数の計算が困難である．また，L2 正則化項の重みは一般的に交差確認などによって複数の候補となる値の中から決定されるが，値を変えて繰り返し計算する際に時間がかかることや，アルゴリズム中で選ばれる特徴量の組合せごとに有効な値が異なるという問題が考えられる．

本研究では，まず一般化線形モデルの特徴選択に対して有効な確率的離散一次法を提案し，L2 正則化項の重みをアルゴリズム中で自動的に決定するための改良を加える．そして，提案手法の有効性を検証するため，数値実験を行う．

Stochastic discrete first-order algorithm for feature subset selection in generalized linear models
Kota KUDO† and Yuichi TAKANO‡
†Degree Programs in Systems and Information Engineering, University of Tsukuba
‡Faculty of Engineering, Information and Systems, University of Tsukuba

2 確率的離散一次法

特徴選択問題は以下のように定式化される．

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & f_\lambda(\boldsymbol{\beta}) \coloneqq f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_0 \leq k. \end{aligned}$$

ここで，$\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ は一般化線形モデルの偏回帰係数のベクトル，k は選択する特徴量の数を表すパラメータ，λ は L2 正則化項の重みを表すパラメータ，$f(\boldsymbol{\beta})$ は一般化線形モデルの損失関数であり，$\|\boldsymbol{\beta}\|_0$ は $\boldsymbol{\beta}$ の非ゼロ成分の数を表す．

目的関数の勾配 $\nabla f_\lambda(\boldsymbol{\beta})$ に対するリプシッツ定数を L とすると，全ての $\boldsymbol{\eta}, \boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ に対して，以下の不等式が成立する．

$$\begin{aligned} f_\lambda(\boldsymbol{\eta}) &\leq \mathcal{Q}_L(\boldsymbol{\eta} \mid \boldsymbol{\beta}) \\ &\coloneqq f_\lambda(\boldsymbol{\beta}) + \nabla f_\lambda(\boldsymbol{\beta})^\top (\boldsymbol{\eta} - \boldsymbol{\beta}) + \frac{L}{2}\|\boldsymbol{\eta} - \boldsymbol{\beta}\|_2^2 . \end{aligned}$$

離散一次法 [1] では，m 回反復時の解 $\boldsymbol{\beta}^{(m)} \in \mathbb{R}^{p\times 1}$ を更新するために，制約条件 $\|\boldsymbol{\eta}\|_0 \leq k$ の下で $Q_L(\boldsymbol{\eta} \mid \boldsymbol{\beta}^{(m)})$ が最小となる $\boldsymbol{\eta}$ を以下のように求める．

$$\begin{aligned} \boldsymbol{\beta}^{(m+1)} &\in \operatorname*{argmin}_{\boldsymbol{\eta}} \left\{ \mathcal{Q}_L(\boldsymbol{\eta} \mid \boldsymbol{\beta}^{(m)}) \;\middle|\; \|\boldsymbol{\eta}\|_0 \leq k \right\} \\ &= \mathcal{H}_k\left(\boldsymbol{\beta}^{(m)} - \frac{1}{L}\nabla f_\lambda(\boldsymbol{\beta}^{(m)}) \right). \end{aligned}$$

ただし，$\mathcal{H}_k(\cdot)$ は絶対値の降順で $k+1$ 番目以降の成分をゼロに変換する作用素を表す．

確率的離散一次法 [2] では，局所的最適解を抜け出して広範囲の解を探索するために，探索点列に確率変動を加える．m 回反復時の解から勾配降下方向に進んだベクトルを $\boldsymbol{c} \coloneqq \boldsymbol{\beta}^{(m)} - (1/L)\nabla f_\lambda(\boldsymbol{\beta}^{(m)})$ としたとき，平均ゼロ，標準偏差 $\sigma^{(m)}$ の独立な正規乱数からなるベクトル $\boldsymbol{\xi} \in \mathbb{R}^{p\times 1}$ を用いて，以下の手順で $\boldsymbol{\beta}^{(m+1)}$ を決定する．

1. $\boldsymbol{c} + \boldsymbol{\xi}$ の成分の置換 π を以下の条件を満たすように定める．

$$\begin{aligned} |c_{\pi(1)} + \xi_{\pi(1)}| \geq |c_{\pi(2)} + \xi_{\pi(2)}| \geq \cdots \\ \geq |c_{\pi(p)} + \xi_{\pi(p)}|. \end{aligned}$$

2. 以下の代入を $j = 1, 2, \ldots, p$ に対して行い，$\boldsymbol{c}$ の k 個の成分を $\boldsymbol{\beta}^{(m+1)}$ に割り当てる．

$$\beta_j^{(m+1)} \coloneqq \begin{cases} c_j & \text{if} \quad j \in \{\pi(1), \pi(2), \ldots, \pi(k)\}, \\ 0 & \text{otherwise.} \end{cases}$$

上記の操作を作用素 $\mathcal{G}_k(\boldsymbol{c} \mid \boldsymbol{\xi})$ とし，解の更新を以下のように表す．

$$\boldsymbol{\beta}^{(m+1)} \in \mathcal{G}_k \left(\boldsymbol{\beta}^{(m)} - \frac{1}{L} \nabla f_\lambda(\boldsymbol{\beta}^{(m)}) \,\middle|\, \boldsymbol{\xi}^{(m)} \right).$$

確率的離散一次法のアルゴリズムを以下に記載する．線形回帰モデルの場合，損失関数 $f(\boldsymbol{\beta})$ には残差二乗和 $(1/2)\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ がよく用いられる．ただし $\boldsymbol{y} \in \mathbb{R}^{p \times 1}$ は応答値のベクトル，$\boldsymbol{X} \in \mathbb{R}^{n \times p}$ は特徴量の行列である．このとき，目的関数の勾配に対するリプシッツ定数は $L = \lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X}) + 2\lambda$ となる．ただし，$\lambda_{\max}(\cdot)$ は行列の最大固有値を表す．また，標準偏差 $\sigma^{(m)}$ は解析的に適切な値を求めることができる [2]．

アルゴリズム 1　確率的離散一次法

ステップ 0：$\boldsymbol{\beta}^{(1)} \in \mathbb{R}^{p \times 1}$ $(\|\boldsymbol{\beta}^{(1)}\|_0 \leq k)$ を決定する．$\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{(1)}$, $m \leftarrow 1$ とする．

ステップ 1：$\sigma^{(m)}$ を決定し，$\boldsymbol{\xi}^{(m)} \sim \mathrm{N}(\boldsymbol{0}, (\sigma^{(m)})^2 \boldsymbol{I})$ によって正規乱数を生成する．

ステップ 2：解を更新する．

$$\boldsymbol{\beta}^{(m+1)} \in \mathcal{G}_k \left(\boldsymbol{\beta}^{(m)} - \frac{1}{L} \nabla f_\lambda(\boldsymbol{\beta}^{(m)}) \,\middle|\, \boldsymbol{\xi}^{(m)} \right).$$

ステップ 3：終了条件を満たす場合，アルゴリズムを停止し，$f_\lambda(\boldsymbol{\beta}^{(m')})$, $m' = 1, 2, \ldots, m+1$ が最小となる $\boldsymbol{\beta}^{(\hat{m})}$ を出力する．

ステップ 4：$m \leftarrow m + 1$ としてステップ 1 に戻る．

3　提案手法

3.1　上界関数の調節

一般化線形モデルで目的関数によく用いられる負の対数尤度の勾配に対して，リプシッツ定数を計算することは困難であり，上界関数 $Q_L(\boldsymbol{\eta} \mid \boldsymbol{\beta})$ を作成できない．そこで，上界関数を自動調節する方法 [3] を採用する．

解を更新するため，リプシッツ定数 L の代わりに $\bar{L} > 0$ を用いて，以下のように $\boldsymbol{\eta}$ を計算する．

$$\boldsymbol{\eta} \in \mathcal{H}_k \left(\boldsymbol{\beta}^{(m)} - \frac{1}{\bar{L}} \nabla f_\lambda(\boldsymbol{\beta}^{(m)}) \right). \tag{1}$$

$f_\lambda(\boldsymbol{\eta}) \geq \mathcal{Q}_{\bar{L}}(\boldsymbol{\eta} \mid \boldsymbol{\beta}^{(m)})$ であれば，定数 $\zeta > 1$ を用いて $\bar{L} \leftarrow \zeta \bar{L}$ とし，式 (1) の計算を繰り返す．そして $f_\lambda(\boldsymbol{\eta}) \leq \mathcal{Q}_{\bar{L}}(\boldsymbol{\eta} \mid \boldsymbol{\beta}^{(m)})$ が成立するとき $\boldsymbol{\beta}^{(m+1)} \coloneqq \boldsymbol{\eta}$ とする．このとき，以下の不等式が成立するため目的関数値が増加しないことを保証できる．

$$\begin{aligned} f_\lambda(\boldsymbol{\beta}^{(m+1)}) &\leq Q_{\bar{L}}(\boldsymbol{\beta}^{(m+1)} \mid \boldsymbol{\beta}^{(m)}) \\ &\leq Q_{\bar{L}}(\boldsymbol{\beta}^{(m)} \mid \boldsymbol{\beta}^{(m)}) \\ &= f_\lambda(\boldsymbol{\beta}^{(m)}) \end{aligned}$$

3.2　L2 正則化項の重みの調節

計算時間の高速化や，適切な L2 正則化項の重みの設定をするために，正則化パラメータを自動調節する方法 [4] を採用する．

パラメータ調節用のデータセットに対する損失関数を $g(\boldsymbol{\beta})$ とする．$g(\boldsymbol{\beta})$ には残差二乗和や負の対数尤度が用いられる．そしてパラメータ調節用のデータセットに対する当てはまりがよくなるように，m 回反復時の L2 正則化項の重み $\lambda^{(m)}$ は以下の式で更新される．

$$\lambda^{(m+1)} = \lambda^{(m)} - \alpha \partial_{\lambda^{(m)}} g(\boldsymbol{\beta}^{(m+1)})$$

ただし，α は移動幅である．$\boldsymbol{\beta}^{(m+1)}$ は $\lambda^{(m)}$ に依存するため，$\partial_{\lambda^{(m)}} g(\boldsymbol{\beta}^{(m+1)})$ を計算することができる．

この方法では，複数の λ の値に対してアルゴリズムを繰り返し最初から実行する必要が無いため計算時間を短縮できる．また，選択される特徴量の組合せごとに λ が調節されるため予測性能の向上が期待される．

4　数値実験

特徴選択問題を解くための既存の発見的解法と提案手法の比較を行い，提案手法の有効性を検証する．提案手法のアルゴリズムや数値実験の詳細は当日述べる．

参考文献

[1] Bertsimas, D., King, A. and Mazumder, R.: Best subset selection via a modern optimization lens. *The Annals of Statistics*, Vol.44, No.2, pp.813-852 (2016).

[2] Kudo, K., Takano, Y. and Nomura, R.: Stochastic discrete first-order algorithm for feature subset selection. *IEICE Transactions on Information and Systems*, Vol.E103-D, No.7, pp.1693-1702 (2020).

[3] Beck, A. and Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, Vol.2, No.1, pp.183-202 (2009).

[4] Luketina, J., Berglund, M., Greff, K. and Raiko, T.: Scalable gradient-based tuning of continuous regularization hyperparameters. *In the 33rd International Conference on Machine Learning*, pp.2952-2960 (2016).