Iterative SIS を用いた類似特徴量抽出の効率化

真鍋晋一郎 [1*]　鳥井修 [1*]

Shinichiro Manabe and Osamu Torii

[1] キオクシア株式会社
[1] Kioxia Corporation
[2] メモリ技術研究所
[2] Institute of Memory Technology Research and Development

Abstract: The theme of this research is to speed up the extraction process of similar features, which has a particularly high calculation cost among the analysis processes of Big Data using sparse modeling and is a major issue for timely processing. When dealing with extremely large data, it is difficult to process it by sparse modeling alone, so screening is performed in advance to reduce the data to some extent. We proposed a method to speed up the extraction process of similar features using this characteristic mechanism, and showed the effect by simulation with artificial data. As the amount of data that can be used increases, its effective use will become even more important.

1　はじめに

計算機システムの高度化と計測技術の急速な進展により，大規模かつ複雑なデータが生み出され，電子的に収集・蓄積されている．データは一般的にデータベースの形で管理されており，これを抽出・分析することで容易には判断できない物事の背後にあるメカニズムや情報パターンを推論することで，社会に役立てる研究が進められている．

特に，製造業においても多くの情報が集められ，ビッグデータとして蓄積されている．これらのデータを AI・機械学習を用いて分析することで，生産性の向上や異常検知，予知保全の精度向上，シミュレーションによる作業の効率化などの取り組みが始まっている．取得可能なデータは種類，量ともに日々増加傾向にある一方で，分析はより短期間に行うことで，その結果をタイムリーに利用したいという要求がある．

データ数に比べて説明変数の数が多いビッグデータでは，目的変数に強い影響を与えている説明変数である特徴量の抽出に，スパースモデリングアルゴリズムの一つである Lasso[1]がよく用いられる．またビッグデータを扱う際, Lasso を単独で適用すると精度や計算時間の観点から困難なことが多く，スクリーニング手法の一つである Iterative SIS[2]を用いて，事前にある程度データを削減する方法もよく用いられている．

Lasso では最終的に得られた特徴量と多重共線性の関係にある説明変数が選択されないことが知られているが，このような説明変数が実用上の観点から重要であることも多い．そのため，特に特徴量と直接相関を持つ説明変数である類似特徴量を，改めてビッグデータから得る手法が提案されている[6]．しかし，この手法はビッグデータに対して特徴量の数に比例した演算を行う必要があり，特に計算に必要なコストが大きい処理の一つとなっている．

そこで本報告では，Iterative SIS の結果を利用し，類似特徴量の抽出処理を高速化する手法を提案する．本手法を用いることで，類似特徴量抽出処理にかかる計算処理が短縮されること，また抽出される類似特徴量自体は従来通りのものであることを人工データを用いたシミュレーションで示す．

本報告の構成は以下の通りである．2 章で問題設定と，提案手法のベースラインとする従来方式について述べる．3 章では提案手法について述べる．4 章では従来手法と提案手法の性能比較のために行った人工データによるシミュレーションについて述べる．5 章では本報告の結論について述べる．

2　問題設定

本章では，前章で述べたLassoの概要，特徴量及び類似特徴量について述べる．またスクリーニング手法の一つであるIterative SISの概要について述べ

*　Shinichiro Manabe(shinichiro.manabe@kioxia.com)
*　Osamu Torii(osamu.torii@kioxia.com)

る．これらから構成される従来手法の全体フローをまとめ，問題設定を一般化する．

2.1 Lassoの概要

近年，情報処理の分野では対象のスパース性を用いたモデリング方法が盛んに提案されている．これは与えられたデータに応じて，そのモデルを説明するために必要な説明変数，すなわち特徴量を自動的に抽出する技術である．与えられたデータが巨大でも，実際の特徴量は少数であるという仮定に基いて変数選択が行われる．こうした手法はスパースモデリングと呼ばれており，その中でLasso (Least Absolute Shrinkage and Selection)と呼ばれる手法が現在広く利用されている．

Lassoとは，変数選択と正則化を伴う回帰分析手法であり，Tibshiraniによって1996年に提唱された[1]．Lassoでは，以下の最適解 (β_0, β) が推定される．

$$\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x_i}^T\boldsymbol{\beta})^2 \quad \cdots (1)$$

$$s.t. \sum_{j=1}^{p}|\boldsymbol{\beta}_j| \le t \quad \cdots (2)$$

ただし，$\boldsymbol{x_i} = [x_{i1}, \cdots, x_{ip}]$: 説明変数，$y_i$: 目的変数であり，n: サンプル数(データ数)，p: 次元数(説明変数の個数)としている．また，t はシュリンケージファクタと呼ばれるパラメータである．式から明らかなように，t→∞ のとき，回帰係数 $(\beta_0, \boldsymbol{\beta})$ は通常の重回帰と一致し，t=0のとき，回帰係数 β はすべて0となり，切片のみのモデルとなる．式(1), (2)は，ラグランジアンを用いて，以下のように変形することもできる．

$$\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x_i}^T\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad \cdots (3)$$

λは正則化パラメータと呼ばれ，上述のシュリンケージパラメータtと1対1対応している．

2.2 特徴量，類似特徴量

Lasso では L1 正則化項を導入しており，この性質によりスパース化が行われ少数の特徴量が選択される．しかし，説明変数間に多重共線性がある場合には，選択される特徴量が不安定になることが知られている．これは相関の高い説明変数同士が，互いに影響を及ぼしあっているためである．

そのため Lasso で選択された特徴量に対して，変数選択される前の全ての説明変数と類似性を調べ，選択されなかった特徴量(以降, 類似特徴量と呼ぶ)を抽出することで，この問題に対応することを考える．

特徴量，及び類似特徴量の概念図を Fig. 1 特徴量と類似特徴量に示す．目的変数である Y（青点）に対して，それに影響を与えている説明変数が特徴量で，図では X（薄い青点）で表示されている．またそれぞれの特徴量に対しての類似特徴量がオレンジ色の点として表されてる．類似度をどのように定めるかにもよるが，通常特徴量に対して類似特徴量は複数個あったり，そもそも一つもなかったりする．

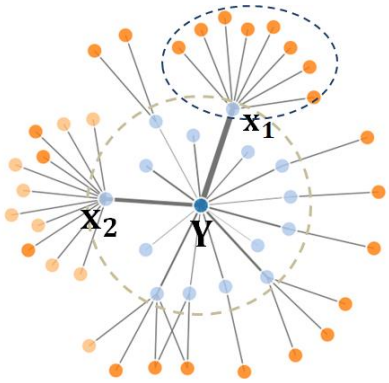

Fig. 1 特徴量と類似特徴量

2.3 Iterative SIS

説明変数の数が非常に多い高次元データを扱う場合，直接 Lassoを用いて特徴量を抽出することは計算量や安定性の観点から一般に困難である．そのため，次元数を事前に落とす必要がある．この処理をスクリーニングと呼ぶ．当然，スクリーニングでは必要な特徴量がデータから失われないことが求められる．これを実現するための手法としてIterative SIS (Iterative Sure Independence Screening) [2]というアルゴリズムが提案されている．

最初にIterative SISで使われているSIS（Sure Independence Screening）について説明する．

求めたい真のモデルを下記のように表現する．

$$M^* = \{1 \le i \le p: \quad \boldsymbol{\beta_i} \ne 0\} \quad \cdots (4)$$

次に下記のωを考える．

$$\omega = X^T \boldsymbol{y},\ \ \omega = (\omega_1, \cdots, \omega_p)^T \quad \cdots (5)$$

このようなωを用いて，元のモデルに対して以下のようなサブモデルを考える．

$$M^{\gamma} = \{1 \le i \le p: \\ |\omega_i| \ is\ among\ the\ first\ [\gamma n] largest\ of\ all\} \quad \cdots (6)$$

ここで$\gamma \in (0,1)$とする．

つまり目的変数と説明変数の内積を目的変数に対する説明力とみなし，内積が大きな説明変数を順に選択する．これはn＜pであるデータを対象とするため，元のp次元のモデルを，γn次元に減らしたものと考えることができる．このようなスクリーニング方法をSIS (Sure Independence Screening)と呼ぶ．

γn はn次元オーダーとなるため，n << pである場合には次元数を大幅に減らす事ができる．しかしSIS には以下のような潜在的な問題があることが知られている．

- 重要な説明変数に強い相関を持つ，重要ではない説明変数が選択される可能性がある
- 説明変数間の多重共線性が変数選択を困難にする

これらの性質を持つ説明変数がビッグデータに存在すると，スクリーニングによって重要な説明変数が保存されない可能性がある．このような問題を回避するためにIterative SISが提案されている．これは反復的に中程度の変数選択を繰り返して，大規模のスクリーニングを行う手法である．Iterative SISではSISで変数選択を行ったあと，Lasso等でモデル選択を行い，その回帰モデルから予測値を得る．そしてその予測値と出力との残差を算出し，これを新しい応答として扱う．また同時に変数選択の対象となるデータから，Lassoで抽出された特徴量を削減することで，次のSISではその前のステップで実施された変数と無相関なものが選択されることになる．この操作は，各ステップごとにSISによって選択される説明変数の優先度を変化させているとも解釈できる．この仕組を用いることで，SISが持っていた潜在的な問題が解消されることが示されている[6].

2.4　全体フローと問題設定

ここまでで Lasso を用いた特徴量の選択，その前段階としての Iterative SIS を用いたスクリーニング，多重共線性の問題に対応するために類似特徴量を選択することを説明した．ここまでの全体のフローを Fig. 2 全体フローに示す．

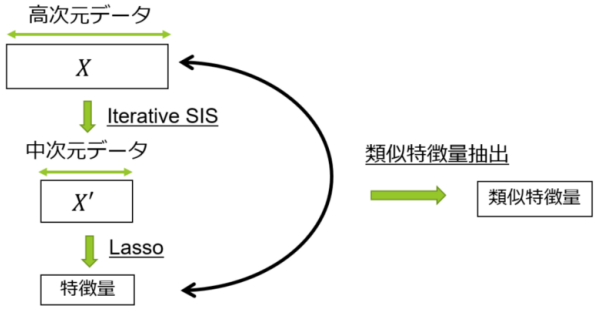

Fig. 2 全体フロー

図からもわかる通り，高次元データであるビッグデータに対して演算する必要がある類似特徴量抽出処理が計算時間上のボトルネックとなっており，この処理を高速化することを目的に設定する．

3　提案手法

Iterative SIS による中間データには SIS 直後のデータである X' と 内部的な Lasso で選択されたデータである X" の 2 種類ある（Fig. 3）．

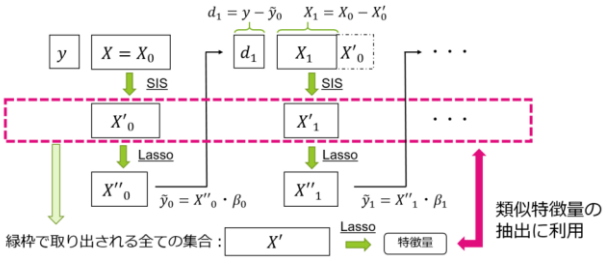

Fig. 3 提案手法

提案手法では，類似特徴量の抽出処理の対象となるデータを元のビッグデータ全体ではなく，X’ に限定する．これは類似特徴量が X’にも残存していると考えられるためである．その根拠は次の通りである．

Iterative SIS は特徴量がデータ内に残るようにスクリーニングをすることを目的としている．よって特徴量は X' にも当然存在している．また SIS による抽出処理では相関係数に似た値を閾値として用いている．つまり特徴量がある SIS により得られたタイミングで，その特徴量に強い相関がある説明変数が同時に取得されているはずである．これはすなわち上記特徴量に対する類似特徴量である．

この考えを Fig. 4 に示す．

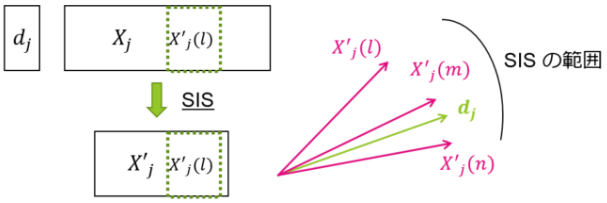

Fig. 4　SIS で選択された各変数の相関イメージ

この性質を利用して類似特徴量の抽出対象を X' に限定することで，類似特徴量の抽出処理にかかる時間の削減を実現する．

4　数値実験

提案手法の有効性を人工データを使ったシミュレーションで検証する．人工データを用いる理由は，各種条件下での効果を定量的に検証するためである．

4.1　実験条件

人工データを下記のような性質を持つように生成する．シミュレーションケースによっては異なる場合もあるため，その際は別途記載する．

- 各説明変数：$x_1, x_2, \cdots, x_p$．ただしpは可変
- 説明変数間の相関：0.3
- 真の特徴量:$x_1, x_2, \cdots, x_5$
- 類似特徴量を持つ特徴量:x_1, x_3
  - x_1 に対応する類似特徴量の数：5
  - x_3 に対応する類似特徴量の数：7
  - 特徴量と類似特徴量の相関閾値：0.7
- データ数：400
- 説明変数の数：500

その他の条件は以下の通り．

- Iteration回数：4
- Iterationごとに一度のSISで得られる説明変数の数：全説明変数の数/20
- 類似特徴量と判断する相関閾値：0.6

4.2　実験結果

提案手法によって得られた特徴量，及び回帰係数を Table 1 に示す．特徴量は，回帰係数が 0 に比べて十分大きな値を持つものを選択した．

Table 1

特徴量	真の回帰係数	回帰係数の推定値
x_1	-3	-2.8633
x_2	1	0.9066
x_3	-2	-1.8620
x_4	6	5.8855
x_5	4	3.8933

適切な特徴量が選択されており，また回帰係数の推定値も真の回帰係数に近い値を持っていることがわかる．

次にそれぞれの特徴量に対して，類似特徴量を取得する．類似特徴量の判断基準は，それぞれの特徴量に対して相関閾値以上の相関をもつ説明変数とする．

Table 2 に，真の類似特徴量の数，従来手法を用いた場合の類似特徴量の数，提案手法を用いた場合の類似特徴量の数，また従来手法と提案手法での計算時間の比率を示す．

Table 2

真の類似特徴量の数	従来手法	提案手法	計算時間の比
12	12	12	3.8333

提案手法でも従来手法同様，全ての真の類似特徴量が抽出できていること，計算時間は従来手法に比べて 3.8 倍程度，高速に行えていることが確認できた．

次にデータ数と説明変数の数の比率による影響を調べるため，データ数を固定したまま，説明変数の数を 100, 200, 300, 400, 500, 1000, 2000, 3000 と変化させてシミュレーションを行った．結果を Fig. 5 に示す．

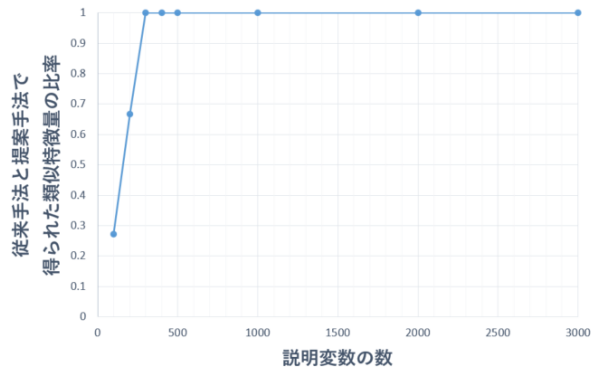

Fig. 5 データ比率による影響

説明変数の数がデータ数に対して特に多い，つまり，よりスパースな構造を持っていても，提案手法でも真の類似特徴量がすべて抽出可能であった．逆に説明変数の数がデータ数に対して少なくなると類似特徴量が抽出できなくなるが，これはIterative SISによるスクリーニングが効きすぎているためである．一般にスクリーニングは説明変数がデータ数に対して十分多いときに減らす目的に利用するものであり，逆に説明変数の数が少ない場合は用いられない．これは重要な説明変数や特徴量も除去されてしまうためである．よって通常の利用では常に有効に機能することが確認できた．

5　終わりに

本報告では，類似特徴量の抽出速度に多大な計算コストがかかっている問題に対して，Iterative SIS の生成データを活用することで類似特徴量の抽出処理を効率化する手法を提案した．

従来手法は，ビッグデータ全てを対象に処理を行うため，データサイズに比例して計算時間が増大していた．数値実験の結果，提案手法は従来手法に比べ 3.83 倍高速に類似特徴量を抽出することができ，また得られた類似特徴量も従来手法と同様のものであることを示した．

規模が大きなデータを扱うほど，従来手法における抽出対象のデータサイズも大きくなり，また存在する特徴量自体も一般的に増加するため，提案手法による効果は飛躍的に高まると考えられる．

参考文献

[1] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.

[2] Fan, Jianqing, and Jinchi Lv. "Sure independence screening for ultrahigh dimensional feature space." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70.5 (2008): 849-911.

[3] Ravikumar, P., Liu, H., Lafferty, J., & Wasserman, L. (2007, December). Spam: Sparse additive models. In Proceedings of the 20th International Conference on Neural Information Processing Systems (pp. 1201-1208). Curran Associates Inc.

[4] Meier, L., Van de Geer, S., & Bühlmann, P. (2009). High-dimensional additive modeling. The Annals of Statistics, 37(6B), 3779-3821.

[5] Yuan, Ming, and Yi Lin. “Model selection and estimation in regression with grouped variables.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1 (2006): 49-67.

[6] 株式会社東芝．解析装置，解析方法，及び，プログラム．特開 2018-151883. 2018-09-27.