ISS-SP-044

潜在意味の重回帰モデルの軽量化

佐藤　将紀†　　三浦　孝夫　††
† 法政大学大学院　理工学研究科　　†† 法政大学理工学部教授

1. はじめに

インターネットの普及により、データ量と種類が膨大になっている。そのため、標本データをすべて得ることは不可能に近い。しかし、データの複雑さを軽減しデータ処理を効率的にすることで、多くのデータが処理可能になる。本論文では、潜在意味解析を適用した重回帰モデルの軽量化(データの縦軸と横軸の縮小)と復元を提案する。説明変数に潜在意味解析を用い、復元したデータから目的変数値を推定する。

2.重回帰モデルの軽量化

説明変数$x_1, \ldots, x_n$および目的変数 Y を使用して分析するデータセット D($x_1, \ldots, x_n$Y)が与えられた場合、有効な合成変数(潜在意味)の D($x'_1, \ldots, x'_n$Y)を取得するために特異値分解を$x_1, \ldots, x_n$に適用する。次に MRA、$f(x'_1, \ldots, x'_n) = a_1 x'_1 + \ldots + a_1 x'_n + b$によって、$x'_1, \ldots, x'_n$上の線形方程式で Y を記述する。D の説明変数に特異値分解を適用し、3 つの行列に分解する($D = U \Sigma V^t$)。説明変数の固有ベクトルとデータ ID の固有ベクトルを抽出する。 累積寄与率を特異値から求める。保持するデータはΣV^tである。特異値の閾値により、保持するデータ量が変わる。閾値以下の特異値を 0 とみなすことでデータ量を減らす。

次に目的変数値の推定を行う。軽量化した潜在意味を復元する。保持するデータΣV^tは、左から U をかけることによって説明変数が復元される。復元した説明変数から重回帰式に当てはめることにより、目的変数値 Y'を推定できる。復元した説明変数から推定した目的変数 Y'と元データの目的変数 Y との誤差を求め、データの削減量と誤差の比較を行い、評価を行う。また、Y'と$x'_1, \ldots, x'_n$の重相関係数により評価を行う。

3.実験

実験では、2000 年から 2003 年までの気象衛星ひまわりから抽出された、日本全国 150 の観測点の 4553 件の気象データを調べる。 16 個の観測変数を分析し、平均気温を目的変数として、他の 15 個の説明変数を次のように設定する。最高温度、最低温度、平均湿度、最小湿度、日照時間、降水量、日降水 mx、降水日数、霧日数、雷日数、真夏日、夏日、熱帯夜、冬日、真冬日。
各次元における目的変数の誤差、Y と Y'誤差の割合とデータ削減量を表にまとめる。

次元数	エラー率(%)	累積寄与率	重相関係数
15	0.111068587	1	0.70732988
14	0.169725341	0.9638803	0.69739824
13	0.180974194	0.9249247	0.68373299
12	0.230723197	0.8797783	0.62987373
11	0.420552408	0.8322311	0.60983723
10	1.366831265	0.7797243	0.58897809
9	2.949136562	0.7253442	0.55725188
8	12.42648781	0.6686186	0.54376826
7	47.39276449	0.6078694	0.52176537
6	52.80229025	0.5413731	0.46387269
5	71.25553842	0.4672778	0.41906383
4	87.75546944	0.3919154	0.31890774
3	113.2304778	0.3093942	0.27893074
2	123.5794875	0.2192201	0.25087937
1	139.2069719	0.0011366	0.23187964

表　各次元にけるエラー率、累積寄与率、重相関係数

結果の意味を説明する。Y と Y'のエラー率の閾値を 1%以下としたときの、データ削減量は約 17%(次元削減数 4)、重相関係数は約 0.61 となった。重相関係数は 0.09 下がったものの、エラー率 1%以下でデータを 17%削減できたことから、潜在意味解析を適用した重回帰モデルは有効的であると考えられる。

4. まとめ

本論文では、潜在気味の重回帰モデルの軽量化を提案した。潜在意味解析により、説明変数の軽量化が可能であることを経験的に示した。本論文によりその有効性を示した。

参考文献

[1] Japan Weather Association: Meteorological Data ″Himawari″,Maruzen 2001
[2] Satoh, M.,and Miura, T.:
Entropy-Based Model Selection Using Monte Carlo Method,Intn'l Conf on Behavioral, Economic, and Socio-Cultural Computing (BESC), China, 2019