v-SVR感度分析法を用いた都道府県間の死亡率格差要因分析

明治大学大学院　先端数理科学研究科　現象数理学専攻　博士前期課程1年
田中　草平

【研究の概要】

- スパース推定において、最も一般的に使われる多重線形回帰モデル(Multiple Linear Regression: MLR)には以下のような問題点がある：
  - 多重共線性
  - 精度の低さ
  - データ数以上の変数を扱えない
- 田辺・鈴木（2020）では機械学習の手法である「サポートベクター回帰（SVR）」を用いて変数選択を可能にする「SVR感度分析法」が提案された。これにより線形モデルの多くの問題を改善することができたが、変数消去のたびにハイパーパラメータ探索の範囲設定を行う必要があり、変数が多い場合に極めて非効率であるという問題点がある。
- 一方で、Smola et al. (2000)ではSVRのハイパーパラメータεの扱いを変えハイパーパラメータの制御をしやすくする"v-SVR"を提案した。
- 本研究では、v-SVRを用いてSVRの感度分析法を行うことで、ハイパーパラメータ探索の範囲設定を1回に限定できる方法を提案する。
- 実証分析では、都道府県間の死亡率を目的変数とし、地域間の健康格差をうむ要因を分析した。
  また、v-SVRを用いた感度分析の有効性の検証のために、通常のSVRの感度分析法（ハイパーパラメータ探索の範囲設定を一回に限定した場合）など他の様々な手法と結果を比較した。

SVRと感度分析法

■ サポートベクター回帰（SVR）

- □ 分類問題の手法であるサポートベクターマシン（SVM）を回帰問題に拡張した手法。
- □ 次のε-不感損失関数を損失関数に用いることによる外れ値に対するロバスト性、カーネル法による非線形性表現能力がある。

ε-不感損失関数：

$$f_\varepsilon\big(y, f(\boldsymbol{x})\big) = \max\{0, |y - f(\boldsymbol{x})| - \varepsilon\}$$

SVRの損失関数：

$$l(\boldsymbol{w}, b) = \frac{1}{2}||\boldsymbol{w}||_2{}^2 + C\sum_{i=1}^{N} f_\varepsilon\big(y_i, f(\boldsymbol{x}_i)\big)$$

※ ε・Cはそれぞれ不感度及び正則化の弱さを表すハイパーパラメータ
※$f(\boldsymbol{x})$との差がε以下のデータはSVMでのサポートベクトル、εはマージン

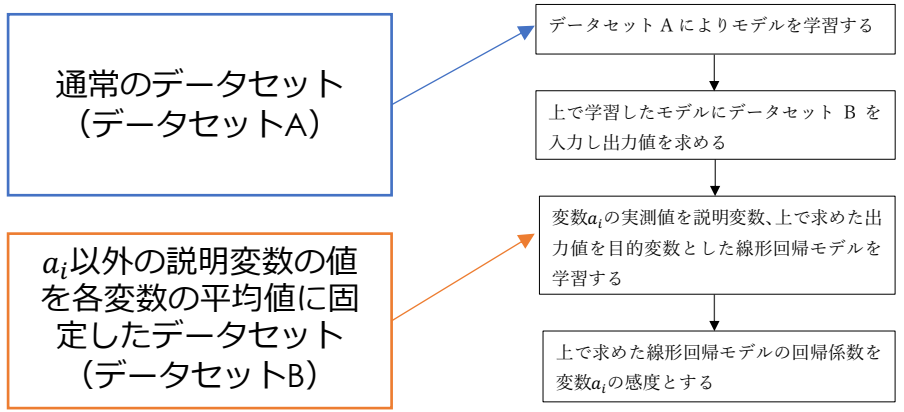
■ SVR感度分析法

- □ SVRは、カーネル法により説明変数を高次元空間へ写すため変数の可読性に欠ける

→ 各説明変数の目的変数への影響度である「感度」を導入し変数選択を可能にする。

- □ 各説明変数の感度を計算し、感度の値が小さいものを順次消去しながらモデルを学習し直すこと繰り返し精度指標（平均二乗誤差等）により最適なモデルを決定する。
- □ 右図のような手順で感度の値を計算し、変数消去・モデルの再学習を繰り返す。

i番目の変数a_iの感度の計算方法



感度分析法の問題点とv-SVR

■ **SVR感度分析法の問題点**
感度分析法は、変数消去のたびにハイパーパラメータ探索の範囲設定を行う必要があり、ハイパーパラメータ探索の範囲の設定を1回に制限した場合、変数消去の過程で最も強く影響するハイパーパラメータεを適切に選択できなくなるという問題点がある。
→ 実行の際極めて非効率

■ **v-SVR(Smola et al.(2000))**
- ハイパーパラメータεを最適化する対象の変数に含めるというアイディアのもと、新たなハイパーパラメータνを導入する。
- νを介して不感度εがモデルの複雑さ(正則化項)と学習データの損失とのトレードオフを制御するような以下の制約付き最適化問題を解くことで回帰関数を求める。

$$\min_{w,\varepsilon,b,\xi_i^+,\xi_i^-} \frac{1}{2}||\boldsymbol{w}||_2{}^2 + C(\nu\varepsilon + \sum_{i=1}^{N}(\xi_i^+ + \xi_i^-))$$

$$s.t.\ \xi_i^- \geq -\big(y - f(\boldsymbol{x})\big) - \varepsilon, \xi_i^- \geq 0, \xi_i^+ \geq y - f(\boldsymbol{x}) - \varepsilon, \xi_i^+ \geq 0, \forall i \in \{1,2,\dots,N\}, \varepsilon > 0$$

- νは、サポートベクトルとなるデータ数の学習データ数に対する割合の下限となる。

→ εをハイパーパラメータとして扱う場合に比べ、νの変動範囲が小さいためグリッドサーチでの範囲の絞り込みがしやすくなる。

実証分析

v-SVRの感度分析法の有効性の検証のために、以下のデータを用いて地域間の健康格差の要因分析を行った。

■ **データ**

- 目的変数
  年齢調整済み平均死亡率
- 説明変数
  健康に影響を及ぼすと考えられる変数として、環境要因28個、生活習慣要因24個
- データの出展
  - 総務省統計局　社会人口統計体系 統計でみる市区町村のすがた
  - 厚生労働省　国民生活基礎調査・国民健康栄養調査・特定健康診査

■ **比較対象**

- 通常のSVRの感度分析法
- 正則化項つきMLR（Ridge、Lasso）
- 部分最小二乗回帰（Partial Least Squared Regression: PLSR）

■ **その他**

MLRを基本とした手法に対して、説明変数間の相関係数が0.7以上のもの同士は目的変数との相関係数が小さい方を消去したデータセットを作り、通常のデータセットと別に適用した（多重共線性回避のための方法の１つ）。

説明変数

環境要因	
核家族	交通事故
高齢単身	公害
婚姻	保健師
収入	社会福祉費
収入格差	児童福祉施設
世帯教育費	高齢就業
自治体教育費	擁護老人ホーム
人口集中	有料老人ホーム
労働災害	介護療養型医療施設
自殺	老人福祉センター
病床	介護老人福祉施設
精神病床	高齢者学級

生活習慣要因	
喫煙	がん検診
飲酒	健康無関心
スポーツ	精神病床
娯楽・趣味	野菜摂取
睡眠時間	食塩摂取
歩数	規則正しい食事
肥満	バランスの取れた食事
メタボ	薄味の食事
健診	食べすぎない意識

結果と考察・まとめ

■ 結果と考察

□ 選択された変数の傾向の違いについて

SVR及びLassoを用いた手法について選ばれた変数の内容を見ると、選ぶ変数の傾向に明確な違いがあることがわかった。選ばれた変数のうち環境要因の変数の数は、以下のような結果となった。

	Lasso	Lasso(変数削減後)	SVR	v-SVR
(環境要因)/(全体)	14/24	9/18	7/16	4/9

MLRは環境要因の変数を重視する傾向があるのに対しSVRは逆に生活習慣要因の変数を重視する傾向があることがわかった。

□ 変数影響度指標の符号及び大きさについて

v-SVR、SVR以外の手法では、回帰係数の符号に明らかな違和感があるものが多く出現した。変数がここまで多くなると、正則化や変数の事前選択等の手法を用いても多重共線性の排除には限界があることを示していると考えられる。SVRではMLRに比べ感度の符号の逆転はかなり排除できており、感度の寄与率の順位もv-SVRの方がより自然な結果となった。

□ 予測精度について

１個抜き交差検証法を用いたMSE（平均２乗誤差）の値は、下表のようになった。予測性精度の観点からも、v-SVRが最も有効であるという結果を得た。

v-SVR	SVR	Ridge	Ridge(変数削減後)	Lasso	Lasso(変数削減後)	PLSR
9.87×10^{-9}	1.32×10^{-8}	4.51×10^{-8}	3.12×10^{-8}	1.31×10^{-8}	1.28×10^{-8}	1.77×10^{-8}

■ まとめ

□ 本研究では、先行研究のSVRの感度分析法において、v-SVRを導入しハイパーパラメータ探索の非効率性を改善した。都道府県別の健康格差の要因分析問題に適用し、この問題に対しての有効性を示した。

□ 今後の展望としては、財務諸表データを用いた信用リスク管理や、自動車事故データを用いた事故リスク要因の推定等に役立てることを検討したい。

選択された変数及び感度と感度の寄与率の値

v-SVR	感度($\times10^{-5}$)	寄与率
健康無関心	7.2440289	18.54%
喫煙	4.3226753	11.07%
睡眠時間	2.3783494	6.09%
肥満	1.5339477	3.93%
がん検診	-0.9439127	2.42%
歩数	-1.1063228	2.83%
婚姻	-1.4887154	3.81%
収入	-2.2524456	5.77%
病床	-2.3942161	6.13%
趣味・娯楽	-2.5514955	6.53%
精神病床	-3.4218699	8.76%
バランスのとれた食事	-4.1748057	10.69%
スポーツ	-5.2509052	13.44%

SVR	感度($\times10^{-4}$)	寄与率
社会福祉費	2.42355	14.83%
がん検診	2.00803	12.29%
健康無関心	1.30650	8.00%
歩数	1.02463	6.27%
睡眠時間	0.93983	5.75%
喫煙	0.92079	5.64%
婚姻	0.50913	3.12%
病床	0.27796	1.70%
精神病床	0.16464	1.01%
肥満	0.06036	0.37%
自治体教育費	0.05386	0.33%
公害	-0.57174	3.50%
スポーツ	-1.47317	9.02%
バランスのとれた食事	-1.48733	9.10%
趣味・娯楽	-1.55594	9.52%
介護老人福祉施設	-1.56223	9.56%