重決定係数を用いたサポートベクトル回帰の実効性評価

垣上　南帆†　　三浦　孝夫†

† 法政大学大学院 理工学研究科 システム理工学専攻　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †naho.kakigami.6p@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

1 前　書　き

サポートベクトル回帰 (SVR) とはサポートベクトルマシン (SVM) の原理を回帰分析に応用した回帰手法である。従来の重回帰分析手法では、多変数線型方程式のうち学習データと最も誤差が少ないものを選んで近似する。一方、SVR では学習データをできるだけ小さな線形区間 (マージン) で近似するが、学習データの次元数を強制的に増加させ、この対応を容易にするカーネル手法も知られる。形式的には高次元学習データを線形近似するため、たとえ元データが非線形であってもこれを線形回帰しているように見える。

つまり、SVR を直接評価する方法は存在しない。カーネル法も入力の次元数を強制的に上げるのみで、非線形回帰といっても実際は高次元空間上で線形回帰をしている。本稿では、SVR が既存の線形回帰分析より回帰精度が優れているかどうかを、主立った回帰手法である線形回帰分析と比較して評価する。

本稿の構成は以下の通りである。第二章では、最小二乗法の線形回帰分析について、さらに第三章では SVM と SVR、カーネル法、リッジ回帰について本研究で必要な内容を述べる。第四章で提案手法について述べ、第五章で実験の準備、結果、考察を示し、第六章で結論とする。

2 最小二乗法線形回帰分析

回帰分析はデータ間の関連を推定する。例に、体重 x_i 身長 x_j から BMIy など説明変数 x で目的変数 y の値を予測する。単回帰分析は説明変数が一つ、重回帰分析は説明変数を複数扱う。線形回帰分析は予測値を説明変数で表す回帰直線、$y = wx + b$ パラメタを推定するものである。線形回帰分析の一つに、次式の誤差の二乗の合計が最小になるようパラメタを決定する誤差最小二乗法がある。

$$\sum_{i=1}^{n}(y_i - f(x_i))^2$$

回帰結果の当てはまりの良さ (評価基準) として重決定係数 R^2 を用いる。これは予測値 f と実測値 y の相関係数 r_{fy} の二乗値である。

$$r_{fy} = \frac{S_{fy}}{S_f S_y}$$
$$R^2 = r_{fy}^2$$

本研究は実効性評価に重決定係数を用いる。

2.1 サポートベクトルマシンとサポートベクトル回帰

SVR とは、 SVM の原理を回帰分析に応用した回帰手法である。

SVM は 2 値分類問題のための機械学習モデルであり、できるだけ大きな区画 (マージン) を分類領域とする線形分類手法である。またカーネル法を利用して非線形データを線形分類するため、許容誤差を柔軟に設定できる。事例から成る学習データ集合 D は、特徴ベクトル x と $-1, +1$ の値を取るクラスラベル y のラベル付きデータで構成されているとする。

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_D, y_D)\}$$

これをマージン最大化によって分離を行う。マージンは分離平面から最近のデータと分離平面との距離で与えられる。検査データが分離直線から離れているほどそのデータを正確に分類できる。また分離平面からの距離がマージン境界上の学習データをサポートベクトルと呼ぶ。分離平面を以下の式とする。

$$y = w \cdot x + b$$

パラメタ w、b、マージン M は点と平面の距離の公式からマージン M は以下の式である。

$$M = \frac{|w^T x_i + b|}{\|w\|}$$

$$\frac{(y_i(w^T x_{i+b}))}{\|w\|} \geqq M, i = 1, 2, 3 \cdots n$$

上記のマージンを最大化するため、Lagrange 法を用いる。両辺を M で割り $\frac{w}{M\|w\|}$ と $\frac{b}{M\|w\|}$ をそれぞれ $\tilde{w}$,$\tilde{b}$ とすると次式で表せる。

$$y_i(\tilde{w}^T x_i + \tilde{b}) \geqq 1$$

ここで扱う最適化問題は次式で定義される。

$$\max_{\tilde{w},\tilde{b}} \frac{1}{\|\tilde{w}\|}$$

$$s.t.\ t_i(\tilde{w}^T x_i + \tilde{b}) \geqq 1, i \in [n]$$

この最大化は $\|w\|$ を最小化すること等しく、今後の計算のために二乗をして最適化問題は次式で表せる。

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t. \forall i y_i(w \cdot x_i + b) \geqq 1$$

SVM アルゴリズムのうちハードマージン SVM では、どの学習データもマージン内に存在できない制約を有する。一方、ソフトマージン SVM ではマージンを超えた距離 ξ_i だけペナルティを与える。ソフトマージン SVM の最小化問題は次式で表せる。

$$\min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t.\forall y_i(w\cdot x_i + b) \geqq 1 - \xi_i$$

スラックペナルティの C は大きいほど誤分類データが与える影響は大きくなる。

SVR では、すべての学習データをマージンに含み、最小になるようなものを選ぶ。SVM と同じく回帰直線からマージン境界上の学習データをサポートベクトルと呼ぶ。最小二乗法の線形回帰分析では回帰直線とデータの距離を誤差としているが、SVR の損失関数ではマージン内であれば誤差は 0 としている。SVR の損失関数は次式で表される。

$$h(y - f(x)) = \max(0, |y - f(x)| - \epsilon)$$

その上で誤差が最小になるようにパラメタを決定するためノイズからの影響を受けにくいという特徴がある。SVR の最適化問題は次式で表せる。

$$(w.b) = \arg\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} h(y_i - f(x_i))$$

$$\epsilon > 0, C > 0$$

以上の最適化問題を解いて得られる SVR の回帰式は次式で表せる。これらの議論は SVM と対をなし、ほぼ SVM と同様の計算技法が使用できる。

$$f(x_i) = \sum_{i=1}^{n}(a_i - a_i^*)K(x_j, x_i) + c$$

$K(x_j, x_i)$ はカーネル関数 K、パラメタと入力の内積であり線形と非線形の形を表現できる。a_i, a_i^* はモデルの複雑さを調整する役割を持ち以下の範囲で値を取る。

$$0 \geqq a_i < C$$

$$0 \geqq a_i^* < C$$

SVM と SVR は共にカーネル関数を使うことで線形・非線形の分離や回帰を可能とする。カーネル法は元の入力データの高次元化を行う。カーネル関数は 2 つの入力から計算される関数 $K(x_j, x_i)$ であり SVR と SVM では入力/高次元化された入力とパラメタの内積はカーネル関数を使って求められる。高次元でデータ数が増えるほどカーネル関数から導かれる分離曲線、回帰式は複雑なものになる。カーネル関数により非線形データも高次元空間へ写像をすることで、SVR では線形回帰が不可能なデータでも回帰が可能になる。求められた回帰式を元データに反映させると非線形に見えるが、実際には高次元空間上でも線形回帰を行っている。カーネル関数には複数の種類があり、本実験では高次元化した際の座標を求めることができる次式の多項式カーネルを用いる。

$$(x^T x^i + 1)^n$$

リッジ回帰は線形回帰分析に重み付けして過学習を防いだ回帰手法である。損失関数は次式で表される。

$$(y - f(x))^T(y - f(x)) + \lambda\|w\|^2$$

$$\lambda > 0$$

リッジ回帰の損失関数では λ で正則化項と損失項でバランスを取っており SVR の損失関数と構造が類似している。λ が大きくなるほど個々のデータからの影響を受けにくくなり、過学習を防ぐ。$\lambda = 0$ で最小二乗法の線形回帰分析と同じであることがわかる。本実験では重回帰分析と SVR のもう一つの比較対象としてリッジ回帰を用いる。

3 提案手法

古典的な線形回帰分析は高速で結果も理解しやすく結果の評価基準も明確である。反面、線形モデルによる近似を仮定するため非線形データには当てはまらないことがある。SVR は非線形データを高次元化し、単純な線形モデルによりマージンの許容誤差に吸収したものである。高速で柔軟に回帰できるが、カーネル関数は説明変数に手を加えて高次元化しているため、本質的に線形重回帰の特性を有する。

そこで、本研究では線形重回帰分析と同様に重相関係数を用いた SVR の精度の評価、および非線形データを高次元化して重回帰分析する方法を提案する。非線形データをカーネル関数で高次元化し線形回帰を行い、見かけ上の非線形回帰することで線形回帰同様に重決定係数を求める。線形/非線形 SVR を重決定係数により線形重回帰、リッジ回帰と実効性比較評価する。特に SVR のサポートベクトルから重相関係数を求めるには SVR の回帰式 $f(x_i) = \sum_{i=1}^{n}(a_i - a_i^*)K(x_j, x_i) + c$ から $a_i - a_i^*$ とパラメタと説明変数の内積 $K(x_j, x_i)$ を求める必要がある。しかし、これは線形/非線形の SVR でも同様であり、例の図 1、図 2 に示した通り行う回帰はいずれも線形回帰である。

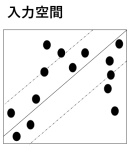

図 1　線 形 回 帰

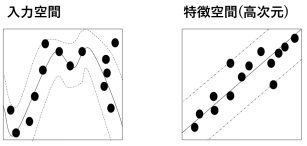

図 2　非線形回帰

4 実　　験

SVM、SVR のパラメタ推定にオープンソースの機械学習ライブラリ LIBSVM を利用した。カーネル関数の種類やハイパーパラメタの設定も細かくすることができる。リッジ回帰はフリーソフト R のパッケージ glmnet で実行。重回帰分析は Excel で行った。高次元化データは多項式カーネル $(x^T x^i + 1)^n$ を用いて計算する。

本実験では気象庁のホームページで公開されている気象データ、千葉県の過去 54 年分の月ごとの気象情報を使用。説明変数 [最高気温],[不照日数] の 2 変数、目的変数 [降水量] 658 件で訓練データは構成される。年, 月データとセットになっており次の表 1 に例として 2020 年のデータを示す。

表 1　2020 年月ごとの千葉県気象データ

年月	降水量の合計 (mm)	最高気温 (℃)	不照日数 (日)
Jan-20	158	17.4	6
Feb-20	35.5	17.8	2
Mar-20	134.5	20.9	6
Apr-20	221	23.1	3
May-20	109	27.7	6
Jun-20	224	32	6
Jul-20	394	31.9	8
Aug-20	54	35.7	0
Sep-20	229	33.5	5
Oct-20	197	26.5	8
Nov-20	15.5	25.1	3
Dec-20	20	16.9	4

データは全て標準化し、線形回帰分析を行う。線形重回帰の回帰直線パラメタは次式から求める。

$$y = ax + b$$

$$a = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

線形 SVR はサポートベクトル数ごとのパラメタを求める。線形リッジ回帰は λ によるパラメタを求める。結果からそれぞれの重決定係数を求める。

非線形回帰では多項式カーネル $(x^T x^i + 1)^n$ による高次元化説明変数を求める。高次元化データを元に線形回帰と同じ手順で線形重回帰、SVR、リッジ回帰を行う。結果からそれぞれの重決定係数を求め、比較をする。

4.1 実 験 結 果

線形重回帰の決定係数を次の表 2 に示す。

表 2　線形重回帰分析 R^2

	決定係数 R^2
線形回帰	0.320143574

線形 SVR はより当てはまりの良い回帰結果を得るようにハイパーパラメタ ϵ,C の値を繰り返し変更し、サポートベクトル数ごとのパラメタを求め、結果を表 3 に示す。

表 3　線形 SVR サポートベクトル数ごとの R^2

SV 数	重決定係数 R^2
23	0.320143095
17	0.320080468
96	0.320069281
29	0.32006482
30	0.320037543
16	0.320030309
14	0.31943303
77	0.319356901
92	0.319163717
5	0.318900657
108	0.318844051
18	0.318500369
8	0.317363703
7	0.317021976
53	0.315430048
49	0.314758339
4	0.308931408
2	0.307587067
3	0.24167292
44	0.232084442
21	0.212052488
21	0.212052488
131	0.002442792

表 3 を見るとサポートベクトル数と重決定係数の大きさは対応していないことがわかる。サポートベクトル数 23 で重決定係数 0.320143095 になり最も大きかった。特に SVR のマージン内に入らないデータの例として 2017 年の 10 月降水量 454 mm 平均気温 28 ℃ 不照日数 12 日、全国的に台風の影響で降水量が増えた月であった。

リッジ回帰は λ の値を 10 倍ずつ大きくして、そのパラメタを求めた。

表 4　リッジ回帰 λ と R^2

λ	重決定係数 R^2
0.0001	0.320143574
0.001	0.320143574
0.01	0.32014356
0.1	0.320142398
1	0.320111054
10	0.320044973
100	0.320028555
1000	0.320026691

λ の値が 0.0001,0.001 のとき重決定係数は線形重回帰分析の結果と同じ値になっている。λ= 0.01 から重決定係数はわずかに下がっていく。SVR のサポートベクトル数 23 の場合と、リッジ回帰 λ= 1 の場合と古典的線形重回帰の結果を次の表 5 に示す。

表 5　線形回帰の R^2

	重決定係数 R^2
SVR(SV 数:23)	0.320143095
線形重回帰	0.320143574
リッジ回帰 ($\lambda = 1$)	0.320111054

線形回帰では回帰手法による重決定係数の差は小さい。

多項式カーネル $(x^T x^i + 1)^n$ による $n = 1$ から $n = 7$ までの多次元説明変数を求めた。

古典的線形重回帰分析の非線形回帰の重決定係数の次元数による推移を次の表 6 に示す。

表 6　非線形線形重回帰 R^2

次元数	線形重回帰 R^2
6	0.328782729
10	0.343382586
15	0.359837661
21	0.378703267
28	0.394808086
36	0.408527438

古典的線形重回帰は説明変数の次元数を上げる度に重決定係数も上がり続けた。

非線形 SVR のパラメタは次元数ごとの結果間で比較ができるようにタ $\epsilon = 0.1$,$C = 1$ に設定。

表 7　非線形 SVR の R^2

次元数	SVR R^2
6	0.320737347
10	0.333639859
15	0.347546899
21	0.356150444
28	0.356632875
36	0.098307689

次元数が 36 まで高次元化すると SVR の決定係数は下がってしまった。また線形回帰同様に SVR はサポートベクトル数によっても決定係数は改善する。次元数 21 でのサポートベクトル数による重決定係数の違いを調べた。

表 8　非線形 SVR サポートベクトル数ごとの R^2

SV 数	重決定係数 R^2
234	0.368404838
232	0.367695459
348	0.365365909
188	0.365161284
297	0.3651136
272	0.363608742
119	0.356897141
564	0.356150444
54	0.307538362
30	0.247967453

サポートベクトル数 234 のとき決定係数は一番大きくなったが、同じ次元数の古典的線形重回帰の決定係数 0.378703267 より約 0.01 低い。

非線形リッジ回帰では SVR と同様の理由で、次元数毎の決定係数がわかるように λ= 1 に設定した。

表 9　非線形リッジ回帰の R^2

次元数	リッジ回帰 R^2
6	0.318386031
10	0.293660687
15	0.296460234
21	0.296061636
28	0.01660777
36	0.315185443

リッジ回帰では次元数の変化で重決定係数が何度か増減を繰り返す。特に次元数 21 から 28 で重決定係数は 0.056 倍になり、その後 28 から 36 で 19 倍になる。

古典的線形重回帰、SVR、リッジ回帰の説明変数の次元数 21 の重決定係数を次の表 10 にまとめた。

表 10　次元数 21 の非線形回帰 R^2

	重決定係数 R^2
SVR(SV 数:564)	0.356150444
線形重回帰	0.378703267
リッジ回帰 ($\lambda = 1$)	0.296061636

次元数 21 まではいずれも重決定係数が大きく減少・増加をすることは無かった。線形重回帰が一番大きい。

古典的線形重回帰、SVR、リッジ回帰の次元数による決定係数の推移を次の表 11 にまとめた。但し、SVR のハイパーパラメタの値設定はそれぞれ $\epsilon = 0.1$,$C = 1$、リッジ回帰は λ= 1。

表 11　次元数ごとの非線形回帰 R^2

次元数	線形重回帰 R^2	リッジ回帰 R^2	SVR R^2
6	0.328782729	0.318386031	0.320737347
10	0.343382586	0.293660687	0.333639859
15	0.359837661	0.296460234	0.347546899
21	0.378703267	0.296061636	0.356150444
28	0.394808086	0.01660777	0.356632875
36	0.408527438	0.315185443	0.098307689

ただし、カーネル関数の変換前データを $\epsilon = 0.1$,$C = 1$ の SVR で回帰するとサポートベクトル数 572、重決定係数 0.215900328。

説明変数の次元数 28 で SVR は一番重決定係数が大きくなったが、次元数 36 で下がってしまう。次元数を上げた精度の向上率では SVR が一番優れている。$\epsilon = 0.1$,$C = 1$ の場合線形 SVR から非線形 SVR で決定係数が 68%増加する。説明変数の次元数 36 での古典的線形重回帰の重決定係数 0.40852743 を SVR が上回ることは無かった。

4.2 考　　察

線形回帰結果では表 5 の通り決定係数はほぼ変わらなかった。重決定係数は重回帰＞ SVR ＞リッジ回帰で重回帰は SVR より 0.50×10^{-6} だけ大きく、わずかに SVR よりも重回帰分析が優れているのではないかと考えられる。

非線形回帰は表 10 の非線形回帰の予測値同士の類似度を、コサイン類似度を用いて計算すると類似度はそれぞれリッジ回帰と重回帰: 0.380359、リッジ回帰と SVR:0.151469、重回帰と SVR:0.935573 だった。また、表 8 の非線形 SVR で最も重決定係数が高かったサポートベクトル数 234 と重回帰の類似度は 0.983912 になっており、重回帰分析との類似度が高いほど重決定係数が高くなる結果になる。

本研究での非線形回帰はいずれも高次元空間上で線形重回帰である。カーネル関数も説明変数の高次元化でのみ使用しており、非線形 SVR にも直接カーネル関数を用いているわけではない。SVR と線形重回帰を最適化法の違う二つの回帰手法だと考えると SVR はパラメタの設定など自由度は高いが、精度の点では線形重回帰が回帰手法としては優秀ではないかと考えられる。

表 7 の通り SVR は次元数が上がると決定係数が最終的に下がってしまったがカーネル関数の選択による影響が考えられる。ただし実験の都合で座標を求める事が可能な多項式カーネルを選択しており、直接座標を求められない RBF カーネルでは決定係数を上げ続けられるがここでは挙げない。

5 結　　論

線形回帰で SVR の重決定係数は重回帰析、リッジ回帰とほぼ変わらずであった。

非線形回帰では線形回帰の結果と比べると SVR は説明変数の次元数が 28 で重決定係数は線形 SVR から決定係数は 65.2%大きくなり次元数 36 で次元数 28 から決定係数は 54.5%下がる。非線形重回帰分析の決定係数は次元数が増えても上がり続け、次元数 36 では線形重回帰分析の決定係数から 27.6%大きくなった。SVR、リッジ回帰の重決定係数が、古典的線形重回帰分析を上回ることは無かった。よって多項式カーネル関数を用いた高次元化座標で非線形回帰を行った場合、重決定係数で評価すると線形重回帰が SVR よりも優れている。

文　　献

[1] Chih-Chung Chang and Chih-Jen Lin”LIBSVM: A Library for Support Vector Machines”2001

[2] 竹内一郎, 烏山昌幸.”機械学習プロフェッショナルシリーズ [サポートベクトルマシン]”講談社,2015

[3] 赤穂昭太郎.”シリーズ 確率と情報の化学 カーネル多変量解析-非線形データ解析の新しい展開”岩波書店,2008

[4] 垣上南帆, 三浦孝夫:サポートベクトル回帰の精度評価, 情報処理学会第 84 回全国大会 (IPSJ),2022, 愛媛