2V-07

サポートベクトル回帰の精度評価

垣上 南帆† 三浦 孝夫†

法政大学理工学部創生科学科

東京都小金井市梶野町 3-7-2

1. 前書き

サポートベクトル回帰(SVR)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元データが非線形でも線形回帰しているように見える。

そのため SVR を直接評価する方法は存在しない。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。

2. 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を予測する。線形回帰は回帰直線$y = wx + b$のパラメタ推定を行う。回帰結果の当てはまりの良さ(評価基準)に重決定係数R^2を用いる。これは予測値fと実測値yの相関係数の二乗値である。

3. サポートベクトル回帰

SVR はすべての学習データを線形区間(マージン)に含み最小になるものを選ぶ。マージン境界上データをサポートベクトルと呼ぶ。SVR の損失関数 h は次式で表される。

$$h\big(y - f(x)\big) = \max\left(0, |y - f(x)| - \varepsilon\right)$$

マージン内データであれば誤差を 0 とし、その上で誤差最小になるようパラメタ決定するためノイズの影響を受けにくい特徴を持つ。

最適化問題と回帰式は次式で表せる。

$$(w.b) = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}||w||^2 + C\sum_{i\in[n]} h\big(y_i - f(x_i)\big)$$

$$\varepsilon > 0, C > 0$$

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

$K\big(x_j, x_i\big)$はカーネル関数、パラメタと入力の内積である。カーネル法は元の入力データの高次元化を行い、入力とパラメタの内積をカーネル関数Kで求めている。SVM/SVR は入力を高次元空間へ写像をすることで非線形な分離と回帰が可能である。本研究は高次元化した際の座標を求められる多項式カーネル$(x^T x^i + 1)^n$を用いる。

SVR と構造が類似している回帰手法にリッジ回帰があり、損失関数は次式で表される。

$$(y - f(x))^T(y - f(x)) + \lambda||w||^2 \quad \lambda > 0$$

λで正則化項と損失項でバランスを取る。

4.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当てはまり基準も明確である。その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。SVR は非線形データを高次元化し単純な線形モデルで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に過ぎないため本質的に線形重回帰である。

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。

5. 実験

5.1 実験準備

SVR のパラメタ推定にオープンソース機械学習ライブラリ LIBSVM を利用した。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。

訓練データは気象庁が公開する千葉県の月ごとの気象データを使用。説明変数[最高気温],[不照日数]、目的変数[降水量] 658 件で構成される。データは全て標準化し回帰分析を行う。

線形 SVR はサポートベクトル数ごとのパラメタを、線形リッジ回帰はλによるパラメタを求める。多項式カーネルよる高次元化説明変数を元に線形回帰と同じ手順で非線形重回帰、SVR、リッジ回帰を行う。

結果からそれぞれ重決定係数を求め比較する。

5.2 実験結果

線形 SVR はサポートベクトル数 23 の重決定係数が最も大きくなった。

SV数	重決定係数R^2							
23	0.320143095	30	0.320037543	92	0.319163717			
17	0.320080468	16	0.320030309	5	0.318900657			
96	0.320069281	14	0.31943303	108	0.318844051			
29	0.32006482	77	0.319356901	18	0.318500369			

表 1：線形 SVR サポートベクトル数ごとの R^2

マージン外データに降水量 454 mm 平均気温 28℃ 不照日数 12 日など台風のあった月がある。

SVR の最大重決定係数とリッジ回帰λ＝1 の場合と線形重回帰の結果を次の表に示す。

	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(λ=1)	0.320111054

表 2:線形回帰の R^ 2

線形回帰での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非線形回帰を行い、決定係数の推移を次の表にまとめた。但し SVR のマージンεと c の値設定はε=0.1,c=1、リッジ回帰は$\lambda = 1$。変換前データをε=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。

次元数	線形重回帰R^2	リッジ回帰R^2	SVR R^2				
6	0.32878273	0.31838603	0.32073735	21	0.37870327	0.29606164	0.35615044
10	0.34338259	0.29366069	0.33363986	28	0.39480809	0.01660777	0.35663287
15	0.35983766	0.29646023	0.3475469	36	0.40852744	0.31518544	0.09830769

表 3:次元数ごとの非線形回帰 R^2

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が SVR の決定係数は次元数が 36 まで高次元化すると下がってしまった。サポートベクトル数によっても決定係数は改善し、次元数 21 のサポートベクトル数 234 で最大の 0.368404838 だったが、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。精度の向上率では SVR が一番優れており線形 SVR から非線形 SVR で決定係数が 68%増加するが、次元数 36 での線形重回帰の重決定係数 0.40852743 が本実験の最大値であった。

5.3 考察

線形回帰結果では表 2 の通り決定係数はほぼ変わらなかった。重回帰は SVR より 0.50×10^{-6}だけ大きく、わずかに SVR より重回帰分析が優れている。表 3 の非線形回帰はいずれも高次元空間上での線形回帰である。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。

6. 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。

参考文献

[1]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](2015) 竹内一郎・烏山昌幸

[2] [カーネル多変量解析-非線形データ解析の新しい展開](2008) 著者:赤穂昭太郎