多変量データにおけるモデルフリーな予測手法

立教大学大学院社会学研究科　高橋淳一
立教大学社会学部　山口和範

１．はじめに

現在、データマイニングで活用されている予測手法のほとんどが、取得したサンプルからモデルを構築し、そのモデルにより指定された説明変数の値における目的変数の値を予測させている。ここで提案する予測手法は、モデルを構築した上で予測を行う今までの手法とは異なり、昨今のコンピュータ処理能力の急激な向上を利用したモデルに依存しない手法である。

データマイニング手法を適用するような大規模なデータの場合、ある意味での分布に関する一様性や、安定性などに疑問がもたれることが多く、すべてのデータの範囲において適応できるようなモデルを構築することは難しい。非線形なモデルを導入しても解決できる問題ではないであろう。いま、目的が構造分析ではなく、予測をすることだけに絞られるような場合には、モデルフリーな方法を利用することが適切であると考えられる。そこで、本報告ではモデルフリーなその予測手順を提唱するとともに、同手法を用いた場合の予測性能に関する数値例についていくつか紹介する。また最後には実用化のために検討すべき課題についても触れる。

２．モデルフリーな予測手法

２．１　予測手順

アルゴリズムの手順は簡単に次のようになる。

① 目的変数と説明変数からなる行列$\boldsymbol{M}$を用意する。
② 行列$\boldsymbol{M}$における目的変数と各説明変数との関連を表現する行列$\boldsymbol{B}$を設定する。
③ 予測したいケースの説明変数ベクトルと、行列$\boldsymbol{M}$におけるケース一つずつの説明変数ベクトルとの距離を行列$\boldsymbol{B}$で調整しながら計算する。
④ 予測したいケースの説明変数ベクトルとの距離の近いケースを行列$\boldsymbol{M}$の中から任意数（ここでは仮にN個とする）取り出す。
⑤ 選ばれたN個のケースにより目的変数yの推定値$\hat{y}_i$を算出し、それを予測値とする。

２．２　アルゴリズム

学習用として使用する行列 $\boldsymbol{M}$ を次のように表す。

$$\boldsymbol{M}=\begin{pmatrix} y_1 & \boldsymbol{X}_1' \\ y_2 & \boldsymbol{X}_2' \\ \vdots & \vdots \\ y_i & \boldsymbol{X}_i' \\ \vdots & \vdots \\ y_N & \boldsymbol{X}_N' \end{pmatrix}=\begin{pmatrix} y_1 & X_{11} & X_{12} & \cdots & X_{1p} & \cdots & X_{1P} \\ y_2 & X_{21} & X_{22} & \cdots & X_{2p} & \cdots & X_{2P} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_i & X_{i1} & X_{i2} & \cdots & X_{ip} & \cdots & X_{iP} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_N & X_{N1} & X_{N2} & \cdots & X_{Np} & \cdots & X_{NP} \end{pmatrix} \qquad (２．１)$$

一方、予測の対象は P 次元のベクトル $\boldsymbol{x}$ で表される。$\boldsymbol{x}$ に対応する y の値を予測することが目的となる。

$\boldsymbol{B}$ を行列 $\boldsymbol{M}$ における目的変数 y と説明変数 X との関係の深さに応じて $\boldsymbol{x}$ に重みづけするための行列として設定する。この行列は $\boldsymbol{\beta}$ を列ベクトルとすると、

$$\boldsymbol{B}=\begin{pmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \cdots & \boldsymbol{\beta}_p & \cdots & \boldsymbol{\beta}_P \end{pmatrix} \qquad (２．２)$$

と表される。行列 $\boldsymbol{M}$ における行ベクトルの添字の集合を I とする。この集合 I の要素 i のなかで、ベクトル $\boldsymbol{B}'\boldsymbol{X}_i$ とベクトル $\boldsymbol{B}'\boldsymbol{x}$ との距離が近いものを N_r 個取り出し、これをサブセット I_r とする。このサブセットは予測したいケース毎に算出する。

サブセット I_r は d を２つの添字ベクトル間の距離を表すものとして、次のように表される。

$$I_r=\left\{ i\,;\, d_{\boldsymbol{B}'\boldsymbol{x}\,\boldsymbol{B}'\boldsymbol{X}_i} \le N_r-th\ smallest\ \ d_{\boldsymbol{B}'\boldsymbol{x}\,\boldsymbol{B}'\boldsymbol{X}_{i\in I}} \right\} \qquad (２．３)$$

こうして得られた添字の集合 I_r によって限定された y_i の値を用いることにより、y の推定を行う。推定値 $\hat{y}$ は関数 w を用いて次のように表現される。

$$\hat{y}=\sum_{i\in I_r} y_i w(y_i) \qquad (２．４)$$

ここで、関数 w は $\sum_{i\in I_r} w(y_i)=1$　を満たすものとする。

関数 w の候補としてはいくつか考えられるが、シンメトリックな確率分布関数に基づくものを用いるのが望ましい。

２．３　補足

①　クラス予測の場合

予測する目的変数 y が質的変数として J 個（$J \geqq 3$）のクラスを持つものである場合は、次のようなプロセスとなる。行列 $\boldsymbol{M}$ における y_i を各クラス j についてのダミー変数 $y_i^{(j)}$ に分割し、上記の予測プロセスを J 回繰り返す。そして、

$$\hat{y}^{(j_0)} = \max_j \hat{y}^{(j)} \qquad (2.5)$$

となる j_0 が予測されるクラスとなる。

②　行列 $\boldsymbol{B}$ の設定方法

$\boldsymbol{B}$ の設定方法として、説明変数の線形結合を行い、次元を縮約した直交空間を構成する Li（1991）によって提案された層別逆回帰（Sliced Inverse Regression;SIR）の手法が有効であると考えられる。

層別逆回帰（以下ＳＩＲと呼ぶ）では次式のモデルを仮定する。

$$y = f(\boldsymbol{\beta}_1\boldsymbol{x}, \boldsymbol{\beta}_2\boldsymbol{x}, \cdots, \boldsymbol{\beta}_K\boldsymbol{x}, \boldsymbol{\varepsilon})$$

ここで、$\boldsymbol{x}$ は p 次元の説明変数を表す列ベクトル、β_k は未知の行ベクトル、ε は独立な確率変数であり、f は $\boldsymbol{R}^{K+1}$ 上の任意な未知の関数である。

ＳＩＲの手法により、p より小さい K 次元の部分空間（これを effective dimension reduction space ; e.d.r 空間と呼ぶ）を構成する $\beta_1, \beta_2, \cdots, \beta_k$ が求められる。これらを用いて $\boldsymbol{B}$ を次式のように設定できる。

$$\boldsymbol{B} = (\boldsymbol{\beta_1} \quad \boldsymbol{\beta_2} \quad \cdots \quad \boldsymbol{\beta_K} \quad \boldsymbol{0} \quad \cdots \quad \boldsymbol{0})$$

$\boldsymbol{B}$ は $p \times p$ の行列だが、ＳＩＲの手法により次元が縮小されて $K+1$ 列目以降は $\boldsymbol{0}$ ベクトルとなるため、実質上、$\boldsymbol{B}'\boldsymbol{X}_i$ と $\boldsymbol{B}'\boldsymbol{x}$ の距離を K 次元で算出することができるようになる。このように、ＳＩＲの手法によって距離を算出するための空間の次元を縮小することができ、目的変数に影響を及ぼすと考えられる変数のみにより、距離を算出することが可能になる。

ＳＩＲの手順の概略は次の通りである。

ⅰ）$\tilde{\boldsymbol{x}}_i = \sum_{\boldsymbol{xx}}^{-1/2}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ とすることにより、$\boldsymbol{x}$を標準化する。

ⅱ）yを $I_1,\cdots,I_H$の H個の層に分割し、層 hに落ちる y_iの比率を $\hat{p}_h$ とする。

すなわち、$\hat{p}_h = (1/n)\sum_{i=1}^{n}\delta_i(y_i)$ となる。但し、δはh番目の層 I_hに y_iが落ちたかどうかを１，０で判定する関数である。

ⅲ）それぞれの層で $\tilde{\boldsymbol{x}}_i$の平均 $\hat{\boldsymbol{m}}_h = (1/n\hat{p}_h)\sum_{y_i\in I_h}\tilde{\boldsymbol{x}}_i$ を算出する。

ⅳ）重みつき共分散行列 $\hat{V} = \sum_{h=1}^{H}\hat{p}_h\hat{\boldsymbol{m}}_h\hat{\boldsymbol{m}}_h'$ を構成し、その固有値と固有ベクトルを求める。

ⅴ）大きい方から K個の固有ベクトルを $\hat{\boldsymbol{\eta}}_k$とし、$\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\eta}}_k\sum_{xx}^{-1/2}$ が求まる。

３．数値例

性能評価のための数値例の題材として、Breiman,Friedman,Olshen,Stone（1987）で用いられている３例を再現した。なお、数値例１～数値例３までにおける同手法の各パラメータは次のものを用いた。

①　（２．２）式に対応するものとして、

$$\boldsymbol{B}' = \begin{pmatrix} \sqrt{\sum_{j\in C}(\bar{X}_{\cdot 1}^{(j)} - \bar{X}_{\cdot 1})^2} & 0 & \cdots & 0 \\ 0 & \sqrt{\sum_{j\in C}(\bar{X}_{\cdot 2}^{(j)} - \bar{X}_{\cdot 2})^2} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{\sum_{j\in C}(\bar{X}_{\cdot P}^{(j)} - \bar{X}_{\cdot P})^2} \end{pmatrix} \qquad (３．１)$$

を用いた。

ここで、$\bar{X}_{\cdot p} = \frac{1}{N}\sum_{i\in I}X_{ip}$ である。$\bar{X}_{\cdot p}^{(j)}$は $y_i = j$となるサイズ $N^{(j)}$の iの集合を $I^{(j)}$として、$\bar{X}_{\cdot p}^{(j)} = \frac{1}{N^{(j)}}\sum_{i\in I^{(j)}}X_{ip}$ と表されるものである。また、ここでは jは二値をとるものとし、$C = \{0,1\}$である。数値例ではいずれもクラスが３個以上あるので、yをダミー変数に分割し、クラス毎に $\boldsymbol{B}$を設定して、最後に（２．５）式によりクラスを予測した。

② 距離dとして次のものを用いた。

$$d_{\boldsymbol{B'x}\,\boldsymbol{B'X}_i} = \left\| \boldsymbol{B'x} - \boldsymbol{B'X}_i \right\|$$

したがって、（２．３）式は、

$$I_r = \left\{ i ; \left\| \boldsymbol{B'x} - \boldsymbol{B'X}_i \right\| \le N_r \text{ - th smallest } \left\| \boldsymbol{B'x} - \boldsymbol{B'X}_{i \in I} \right\| \right\}$$

として、添字の部分集合を求めた。

③ 関数wとして一様密度関数を用いた。すなわち、$w = \dfrac{1}{N_r}$として、（２．４）式は、

$$\hat{y} = \frac{1}{N_r} \sum_{i \in I_r} y_i$$

として推定を行った。

３．１　数値例１

デジタル時計や計算機等では、図１に示されるように、７つの発光体を用いることにより数字を表現している。

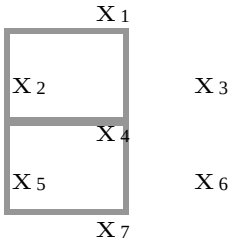

図１．発光体の変数割当

この７つの発光体を１か０かの二値をもつ７つの変数とみなすと、その７つの変数のパターンから１０個のクラスを識別することができる。（図２ & 表１）

図２．クラス（１～０）の７変数による表現

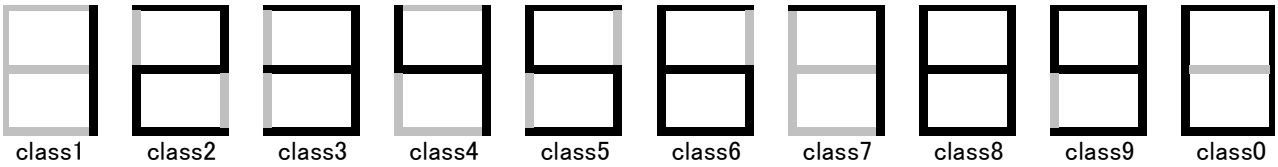

表１．クラス（１～０）の７変数による表現

class	x1	x2	x3	x4	x5	x6	x7
1	0	0	1	0	0	1	0
2	1	0	1	1	1	0	1
3	1	0	1	1	0	1	1
4	0	1	1	1	0	1	0
5	1	1	0	1	0	1	1
6	1	1	0	1	1	1	1
7	1	0	1	0	0	1	0
8	1	1	1	1	1	1	1
9	1	1	1	1	0	1	1
0	1	1	1	0	1	1	1

表１に基づいて、シミュレーション用のデータを生成した。この際、各値に１０％の確率でエラーを生じさせた。したがって、たとえばクラス５のケースが表１に示されているような正しい表示をする確率は、

$$P((x_1,\cdots,x_7)=(1,1,0,1,0,1,1)\,\big|\,Y=5)=(0.9)^7$$

となる。

学習用データのサンプルサイズを１０００、予測するデータのサンプルサイズを１０００とし、提唱手法に加えて、決定木（CART・QUEST）、ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ（ﾊﾞｯｸﾌﾟﾛﾊﾟｹﾞｰｼｮﾝ法）の各手法を用いて予測を行い、その誤分類率Rを比較した。（表２）

表２．数値例１における手法別の誤分類率R

手法	R
決定木（CART）	32.6%
決定木（QUEST）	33.3%
ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ	35.3%
提唱手法	29.6%

なお、R は以下のように算出される。

$$R=1-\frac{1}{N}\sum_{i=1}^{N}\xi\left[\hat{y}_i=y_i\right]$$

ここで、$\xi[\cdot]$は括弧内が正しければ１、誤りならば０をとる関数である。

提唱する手法は他の手法に比較して非常に誤分類率の低い、精度の高い予測をしていることがわかる。

３．２　数値例２

同手法を用いる上でもっとも懸念されるのは目的変数にほとんど影響を及ぼさないノイズ変数が混在した場合に、依然として精度の高い予測が行えるかということがある。この点を検証するために、デジタル時計の拡張版として、７つの変数の他に１７個のノイズ変数を追加したタイプのものを考える。（図２）ノイズ変数は１もしくは０となる確率がともに５０％の、目的変数に対して独立な変数である。
同様に学習用データのサンプルサイズを１０００、予測するデータのサンプルサイズを１０００として、各手法の誤分類率 R を比較した。（表３）

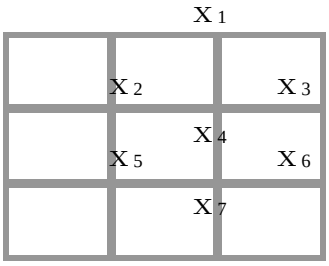

図２．発光体の変数割当

表３．数値例２における各手法の誤分類率 R

手法	R
決定木（CART）	30.5%
決定木（QUEST）	30.5%
ニューラルネットワーク	29.7%
提唱手法	23.3%

ここでも提唱する手法は高い精度を保っている。

３．３　数値例３

続いて、変数が間隔尺度となるケースについてのシミュレーションを行った。図３のような、$h_1(t), h_2(t), h_3(t)$ の３つの波形要素を組み合わせることによって、３つのクラスを構成させる。

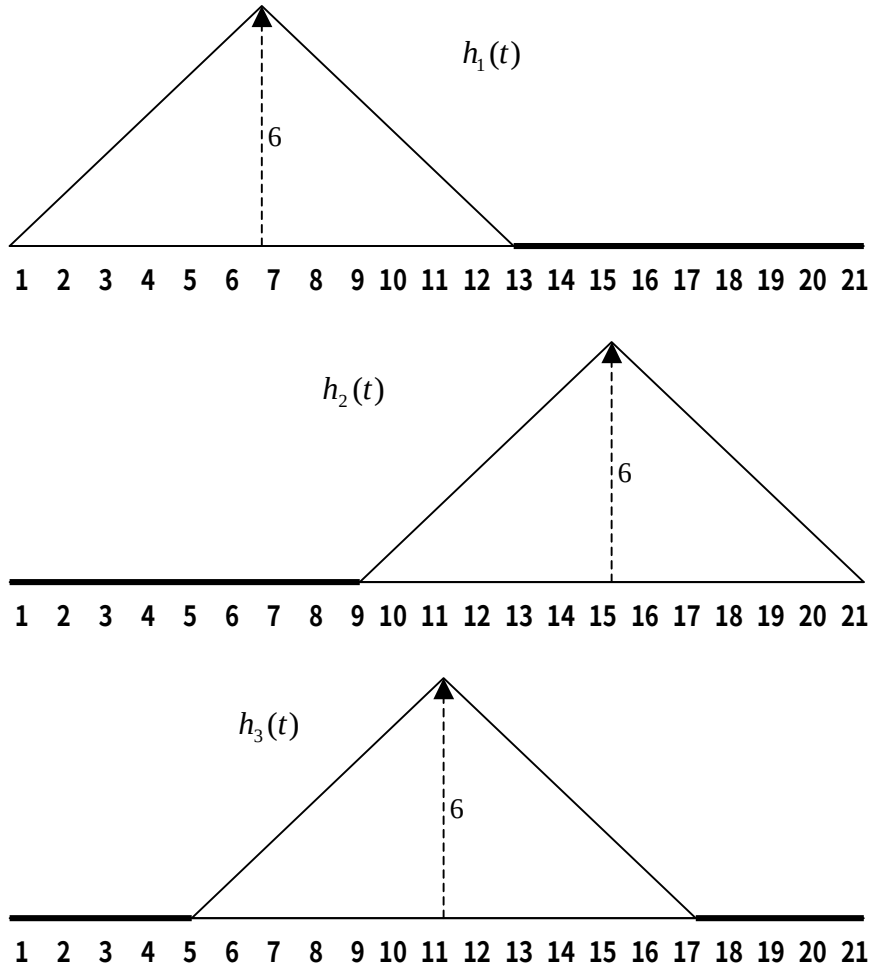

図３．波形要素

それぞれのクラスは次のようにこれらの波形要素のうち２つを組み合わせて構成される。

$$class\,1: x_m = uh_1(m) + (1-u)h_2(m) + \varepsilon_m,\ m = 1,\cdots,21$$
$$class\,2: x_m = uh_1(m) + (1-u)h_3(m) + \varepsilon_m,\ m = 1,\cdots,21$$
$$class\,3: x_m = uh_2(m) + (1-u)h_3(m) + \varepsilon_m,\ m = 1,\cdots,21$$

uは０≦u＜１の範囲内の一様分布、εは平均０、分散１の正規分布に従う確率変数である。学習用データ、予測用データのサンプルサイズはそれぞれ６００とした。各手法の誤分類率Rを比較したものが表４である。

表４．数値例３における各手法の誤分類率R

手法	R
ロジスティック回帰	66.9%
決定木（CART）	18.5%
決定木（QUEST）	17.5%
ニューラルネットワーク	74.8%
提唱手法	10.0%

このように、変数が間隔尺度となる場合においても、提唱する手法は高い精度を示した。

３．４　数値例４

次に$\boldsymbol{B}$の設定にＳＩＲの手法を用いた場合の数値例を示す。次式を用いてデータを生成した。

$$y = x_1(x_1 + x_2 + 1) + 0.5\boldsymbol{\varepsilon}$$

$x_1, x_2, \boldsymbol{\varepsilon}$のそれぞれは平均０、分散１の標準正規分布をする互いに独立な変数である。この他、同様に独立に標準正規分布に従うx_3からx_{10}までの８つの変数を用意した。したがって、Ｐ＝１０となり、サンプルサイズ Ｎ は１０００とした。ＳＩＲの手法によって、βを出力した結果が表５である。

表５．βの各要素

	$\beta 1$	$\beta 2$	$\beta 3$	$\beta 4$
固有値	3.02	0.10	0.01	0.00
固有ベクトル	−0.09	0.94	−0.28	−0.07
	0.90	0.07	0.02	0.03
	0.21	0.17	0.34	−0.11
	0.07	0.05	0.49	−0.05
	−0.14	0.03	0.03	0.84
	−0.25	0.01	0.35	−0.45
	−0.03	0.24	0.62	0.27
	−0.18	0.00	0.10	−0.03
	−0.07	0.10	0.13	0.06
	−0.07	0.04	0.20	−0.01

これらのβから行列 $\boldsymbol{B}$ を構成して予測を行った。学習用データによって学習用データ自身を予測させて *bias* と *MSE* を算出した結果が表６と表７である。表６は *bias* と *MSE* を算出する際、 y_i として、 $y = x_1(x_1 + x_2 + 1) + 0.5\boldsymbol{\varepsilon}$ を用いたもので、表７は誤差項を除いた、 $y = x_1(x_1 + x_2 + 1)$ を用いたものである。いずれにおいても、ケースの制限数 *Nr* とβの次元 *K* を変化させて比較した。

なお、*bias* と *MSE* は以下のように算出した。

$$bias = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) \qquad MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

表６．学習用データに関する bias と MSE（理論値に誤差項を含む場合）

	K=2		K=3		K=4	
Nr	bias	MSE	bias	MSE	bias	MSE
3	−0.01	0.72	−0.06	0.50	−0.10	0.56
5	−0.02	0.87	−0.08	0.65	−0.13	0.74
10	−0.08	1.07	−0.13	0.85	−0.16	0.95
20	−0.11	1.24	−0.17	1.06	−0.20	1.22
50	−0.18	1.50	−0.23	1.45	−0.24	1.65

表７．学習用データに関する bias と MSE（理論値に誤差項を含まない場合）

	K=2		K=3		K=4	
Nr	bias	MSE	bias	MSE	bias	MSE
3	−0.02	0.65	−0.07	0.43	−0.11	0.51
5	−0.03	0.72	−0.09	0.53	−0.14	0.62
10	−0.09	0.89	−0.14	0.67	−0.17	0.79
20	−0.12	1.04	−0.18	0.86	−0.21	1.03
50	−0.19	1.29	−0.24	1.24	−0.25	1.45

次に学習用データとは別にテスト用データを用意して予測をした結果の *bias* と *MSE* を表８に示す。

ここではy_iとしては、 $y = x_1(x_1 + x_2 + 1) + 0.5\varepsilon$　に基づいている。

表８．テスト用データに関する bias と MSE

	K=2		K=3		K=4	
Nr	bias	MSE	bias	MSE	bias	MSE
3	−0.03	1.48	−0.11	1.04	−0.15	1.21
5	−0.03	1.37	−0.14	1.02	−0.17	1.12
10	−0.07	1.31	−0.19	1.07	−0.21	1.17
20	−0.12	1.37	−0.23	1.20	−0.25	1.33
50	−0.19	1.57	−0.29	1.51	−0.31	1.70

比較のため、最小二乗法により同様の数値を出力した。
モデルとして以下の２つをもとに最小二乗法を適用した結果が表９である。

モデルＡ：$y = a_0 + a_1x_1 + a_2x_1^2 + a_3x_1x_2 + a_4x_2^2 + a_5x_2$

モデルＢ：$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_{10}x_{10}$

表９．最小二乗法を適用した場合

	bias	MSE
モデルA	0.00	0.24
モデルB	0.00	3.46

この結果を表７と比較すると、提唱手法による結果はモデルＡによる最小二乗法には及ばないが、モデルＢによるものの精度を大幅に上回っていることがわかる。

４．考察

４．１　距離算出のための空間設定

同手法では$\boldsymbol{B}$を用いて距離を算出するための空間を設定している。ここでの目的は各説明変数が目的変数に及ぼす影響度合いを空間内で表現することである。さらにＳＩＲを用いる場合には、この過程で次元の縮小も行われる。もし、目的変数に全くあるいはほとんど影響を及ぼさない説明変数が存在し、この過程でその影響を取り除くことができなければ、次に行われる距離の算出においてケースの順位づけが適正に行われず、予測精度の著しい低下をまねくこととなる。尚、ＳＩＲを用いた場合、距離の算出には数値例で用いたようなユークリッド距離を使用すればよい。

４．２　N_rの設定

推定に用いられるケース数に制限を加えるための N_rの設定を行う方法としては、クロスバリデーション法を用いるのが有効である。次の式で表される誤分類率の推定値 R^{CV}の値が最小となる N_rを選び出す。

$$R^{CV}\left[\hat{y}(N_r)\right] = 1 - \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\xi}\left[\hat{y}_{-i}(N_r) = y_i\right]$$

ここで、$\boldsymbol{\xi}[\cdot]$は括弧内が正しければ１、誤りならば０をとる関数である。また、$\hat{y}_{-i}(N_r)$は i番目のケースをそれ以外のケースで推定したものである。

ここでは、使用するデータを制限する基準としてケース数を採用しているが、もう一つの制限方法としてケースからの距離で制限するものが考えられる。但し、距離で制限する場合、以下のような問題点がある。

$\boldsymbol{x}$が高次元になればなるほど、あるデータの近傍にはほとんどあるいは全くデータが存在しない可能性が非常に高くなる。したがって、距離で制限を行った場合に、その制限下で用いることができるデータが全く存在ないというケースが高い可能性で考えられる。こういったケースでは局所的に制限する距離を緩めなければならなくなり、その結果、ケース数で制限しているのとアルゴリズム上さほど差異がなくなってくる。こういった理由から今回は制限方法として距離ではなく、ケース数を採用することにした。

また、ＳＩＲの手法を用いた次元の縮小が、こういった高次元におけるデータ不足を解決する一つの手段となり得るものと考えられる。

４．３　wの設定

制限されたケースからどのように推定値を得るか、すなわち（２．４）式における局所的な確率密度関数wをどのような形にするかという点も検討課題の一つである。今回の数値例では単純に y_iの平均をとったが、選ばれたケース $\boldsymbol{X}_i$の中でも $\boldsymbol{x}$に近い値ほど重視したいと考えるならば、関数wにカーネル密度関数などを導入し、距離に応じた荷重つきの平均をとるなどの方法も検討できるであろう。

なお、予測誤差の推定に関する検討が今後必要となる。

この手法の大きな要点は事前に得られている目的変数つきのサンプルを、そこからモデルをつくるのではなく、直接予測をするために活用しようとする点である。コンピュータの性能の飛躍

的な向上により、得られたサンプルからモデルをつくって情報を縮約するのではなく、そのままの形で予測に利用するという方法が可能になってきたと考えられる。今回の数値例でも、学習用サンプル、予測用サンプルのサンプルサイズを各々１０００として予測を行ったが、その際に計算される距離の数は１０００×１０００の１０万件となる。にもかかわらず、コンピュータが算出に要する時間は数秒であった。学習用サンプルのサンプルサイズはコンピュータの処理能力とを加味した上で決定する必要がある。仮りにコンピュータの処理能力をはるかに超える大きなサイズのサンプルを得ることができた場合、予測の際に用いるモデル用サンプルはそこから無作為抽出によりコンピュータの性能に耐え得るサイズにまで縮小されたものを使用することができる。このようにサンプルサイズを調整することにより、予測に際して、コンピュータの性能を効率よく活用することが可能になってくる。

参考文献

[1]Breiman,L.,Friedman,J.H.,Olshen,R.,and Stone,C.J.(1984). Classification and Regression Trees. Belmont,CA:Wadsworth.

[2]Fox,J.(2000). Multiple and Generalized Nonparametric Regression. Papers Series on Quantitative Applications in the Social Sciences, 07-131. Thousand Oaks, CA:Sage.

[3]Li,Ker-Chau(1991). Sliced Inverse Regression for Dimension Reduction. JASA Vol.86 No.411.

[4]Muller,Hans-Georg(1988). Nonparametric regression analysis of longitudinal data. Springer-Verlag.

[5]Simonoff,J.A.(1996). Smoothing Methods in Statistics. New York:Springer-Verlag.