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している。ここで提案する予測手法は、明示的な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、目的が構造分析でなく、予測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。

2. モデルフリーな予測手法

2.1 予測手法

アルゴリズムの手順は簡単で次のようになる。

① 目的変数と説明変数からなる行列\( M \)を用意する。
② 行列\( M \)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\( B \)を層別逆回帰（Sliced Inverse Regression:SIR）(Li 1991)等を用いて設定する。
③ 予測したいケースの説明変数ベクトルと、行列\( M \)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\( B \)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\( M \)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\( \hat{y}_j \)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\( M \)を次のように表す。

\[
M = \begin{pmatrix}
\{y_1 \mid X_1'\} & \{y_2 \mid X_2'\} & \ldots & \{y_M \mid X_M'\}
\end{pmatrix}
\]

（2.1）

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

この行列は\( B \)を列ベクトルとすると、

\[
B = (\beta_1 \beta_2 \ldots \beta_p \ldots \beta_P)
\]

（2.2）
と表される。行列 \( \mathbf{M} \) 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 \( \mathbf{B}'x_i \) とベクトル \( \mathbf{B}'x \) との距離が近いものを \( N \) 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。

\[
I_r = \{ i; d_{\mathbf{B}'x \mathbf{B}'x_i} \leq N_r - th \text{ smallest } d_{\mathbf{B}'x \mathbf{B}'x_i} \} \tag{2.3}
\]

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。

\[
\hat{y} = \sum_{i \in I_r} y_i w(y_i) \tag{2.4}
\]

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

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（\( J \geq 3 \)）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 \( \mathbf{M} \) 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、

\[
\hat{y}^{(j)} = \max_j \hat{y}^{(j)} \tag{2.5}
\]

となる \( j_0 \) が予測されるクラスとなる。

(2) 行列 \( B \) の設定方法

\( B \) の設定方法として、説明変数の線形結合を行い、次元を縮小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（SIR）では一次式のモデルを仮定する。

\[
y = f(\beta x, \beta x \Lambda, \beta x \epsilon)
\]

ここで、\( x \) は \( p \) 次元の説明変数を表す列ベクトル、\( \beta \) は未知の行ベクトル、\( \epsilon \) は独立な確率変数であり、\( f \) は \( R^{k+1} \) 上の任意な未知の関数である。SIR の手法により、\( y \) として小さい \( k \) 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 \( \beta_1, \beta_2, \ldots, \beta_k \) が求められる。これらを用いて \( B \) を一次式のように設定できる。

\[
B = (\beta_1, \beta_2, \Lambda, \beta_k, 0, \Lambda, 0)
\]

\( B \) 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 \( 0 \) ベクトルとなるため、実質上、\( \mathbf{B}'x_i \) と \( \mathbf{B}'x \) 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を選び、距離を算出することが可能になる。SIR の手法の概念は次のようにある。

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

2） \( y \) を \( I_1, \ldots, I_n \) の \( H \) 個の層に分割し、層 \( h \) に落ちる \( y \) の比率を \( \hat{p}_h \) とする。

すなわち、\( \hat{p}_h = (1/n) \sum_{i \in I_h} \delta_i(y_i) \) となる。ただし、\( \delta \) は \( h \) 番目の層 \( I_h \) に \( y \) が落ちたかどうかを \( 1, 0 \) で判定する関数である。
3）それぞれの層で $\bar{x}_i$ の平均 $\bar{m}_k = (l/n_k)\sum y_i x_i$ を算出する。

4）重みつき共分散行列 $\hat{\Sigma} = \sum_{i=1}^{n} \hat{\beta}_k \bar{m}_k \bar{m}_k^t$ を構成し、その固有値と固有ベクトルを求める。

5）大きい方から $K$ 項の固有ベクトルを $\hat{\eta}_k$ とし、$\hat{\beta}_k = \hat{\eta}_k \sum_{k=1}^{n} \hat{\eta}_k^t$ が求まる。

$B$ の設定方法として、SIRの他に Mizuta and Minami (2000) により提案された射影追跡法を利用した SIR pp (SIR with projection pursuit;SIRpp) の利用も考えられる。

3．数値例

数値例として次式を用いてデータを生成した。

$$y = x_1(x_1 + x_2 + 1) + 0.5 \epsilon$$

$x_1, x_2, \epsilon$ のそれぞれを平均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 $x_1$ 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 $\beta$ を出力し、これらの $\beta$ から行列 $B$ を構成して予測を行った。学習用データと比べてテスト用データを用いて予測をした結果の bias と $MSE$ を表1に示す。

<table>
<thead>
<tr>
<th>K=1</th>
<th>K=2</th>
<th>K=3</th>
<th>K=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr</td>
<td>bias</td>
<td>MSE</td>
<td>bias</td>
</tr>
<tr>
<td>3</td>
<td>0.04</td>
<td>1.24</td>
<td>0.07</td>
</tr>
<tr>
<td>5</td>
<td>0.04</td>
<td>1.14</td>
<td>0.08</td>
</tr>
<tr>
<td>10</td>
<td>0.05</td>
<td>1.11</td>
<td>0.11</td>
</tr>
<tr>
<td>20</td>
<td>0.07</td>
<td>1.14</td>
<td>0.13</td>
</tr>
<tr>
<td>50</td>
<td>0.12</td>
<td>1.28</td>
<td>0.20</td>
</tr>
</tbody>
</table>

比較のため、最小二乗法及び決定木により同様の数値を出力した。その結果が表2である。

尚、最小二乗法ではモデルとして以下の2つをもとに bias と $MSE$ を算出した。

モデルA: $y = a_0 + a_1 x_1 + a_2 x_2^2 + a_3 x_1 x_2 + a_4 x_2 + a_5 x_2^2$

モデルB: $y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_2$

表2．最小二乗法及び決定木を適用した場合

<table>
<thead>
<tr>
<th></th>
<th>bias</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>最小二乗法(モデルA)</td>
<td>0.00</td>
<td>0.24</td>
</tr>
<tr>
<td>最小二乗法(モデルB)</td>
<td>0.00</td>
<td>3.46</td>
</tr>
<tr>
<td>決定木(CART)</td>
<td>0.03</td>
<td>1.13</td>
</tr>
</tbody>
</table>

提唱手法による結果は最小二乗法においてはモデルAには及ばないが、モデルBによるもののが精度を大幅に上回っていることがわかる。また、決定木も非常に良い精度で予測を行っているが、最も $N_k$ を選んだ場合の提唱手法はその精度を上回っている。
4. 考察

4.1 距離算出のための空間設定

同手法では$B$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。SIRを用いることによって、この過程で次元の縮小も行われる。もし、目的変数に全くあれば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すこととなる。尚、距離の算出については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。次式で表現される誤分類率の推定値$R^C_P$の値が最小となる$N_r$を選ばる。

\[ R^C_P (N_r) = 1 - \frac{1}{N} \sum_{i=1}^{N} \xi [\hat{y}_i(N_r) = y_i] \]

ここで、$\xi [\cdot]$は括弧内が正であれば1、逆なら-1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。

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。単純に$y$の平均をとるという方法に、選ばれたケース$X_i$の中で$x_i$に近い値ほど重視したいと考えるならば、関数$w$をカーネル密度関数などを導入し、距離に応じた重き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素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。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を含む上で決定する必要がある。仮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る。

参考文献
