プライバシーを保護する線形回帰分析における変数選択法

1X09C121-2　湯川 輝一朗
指導教員　後藤 正幸

1　研究背景・目的

複数のメンバーが分散保持する情報を秘匿したまま知識発見を行うプライバシー保護データマイニング (以下，PPDM) が近年の高度情報化や個人情報保護の観点から注目されている [1]．PPDM では，各メンバーが所持するデータを互いに公開せずにデータマイニングを行うことを目的としている．

PPDM には既に様々な手法が存在しているが，本研究では垂直分割されたデータを用いる回帰分析を対象とする．データの垂直分割とは，各メンバーがサンプルに対して異なる説明変数のデータを所持するものである．この手法の一つに，須子らの手法 [1] がある．須子らの手法は，全ての説明変数を用いてプライバシーを保護しながら回帰係数の最小二乗推定量を分散計算により求める手法である．

しかし，学習データが与えられたもとで未観測データを予測する問題を考えたとき，全ての説明変数を含んだ重回帰モデルが最も予測精度の高いモデルとは限らない．そこで，本研究では AIC 基準を用いて全ての説明変数から目的変数に寄与する説明変数のみをモデルに含むよう変数選択を行い，予測精度の高いモデルを構築するプロトコルを提案する．変数選択では，プライバシー保護の観点から変数減少法を用い，互いのデータを秘匿しながら変数選択を行う手法を示す．また，シミュレーション実験により予測精度の観点から提案プロトコルの有用性を示す．

2　従来手法

Alice，Bob の二者で垂直分割されたデータを用いて回帰分析を行う場合を考える．Alice と Bob はサンプルに対して，異なる説明変数を所持するように分割されており，Alice は p 個，Bob は q 個の説明変数に関するデータを所持しているものとする．いま，サンプル数 N，$p+q$ 個の説明変数がによる説明変数行列 $\boldsymbol{X}$ を，$\boldsymbol{X}=[x_{i,j}]$ とする．但し，$i=1,2,\cdots,N$，$j=1,2,\cdots,p+q$ である．このとき，Alice と Bob は $\boldsymbol{X}$ のうち説明変数行列 $\boldsymbol{X}_{\mathrm{A}}=[x_{i,s}]\in\mathbb{R}^{N\times p}$，$\boldsymbol{X}_{\mathrm{B}}=[x_{i,t}]\in\mathbb{R}^{N\times q}$ をそれぞれ所持するものとする．但し，$s=1,2,\cdots,p$，$t=p+1,p+2,\cdots,p+q$ である．回帰分析を行うにあたり，Alice と Bob は，互いに $\boldsymbol{X}_{\mathrm{A}},\boldsymbol{X}_{\mathrm{B}}$ を公開せず，目的変数データ $\boldsymbol{y}=(y_1,y_2,\cdots,y_N)^T$ を二者で共有するものとする．この時，Alice は，$\boldsymbol{X}_{\mathrm{A}}$ に対応する回帰係数ベクトル $\boldsymbol{\beta}_{\mathrm{A}}\in\mathbb{R}^p$，Bob は，$\boldsymbol{X}_{\mathrm{B}}$ に対応する回帰係数ベクトル $\boldsymbol{\beta}_{\mathrm{B}}\in\mathbb{R}^q$ と定数項 β_0 を最小二乗法を用いて推定する．本研究で扱う重回帰モデルは式 (1) で定義される．

$$\boldsymbol{y}=\mathbf{1}\beta_0+\boldsymbol{X}_{\mathrm{A}}\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{X}_{\mathrm{B}}\boldsymbol{\beta}_{\mathrm{B}}+\boldsymbol{\varepsilon}. \tag{1}$$

ここで，$\mathbf{1}$ は要素が全て 1 の長さ N の列ベクトルであり，$\boldsymbol{\varepsilon}$ は正規誤差ベクトルである．この時，Alice が推定する回帰ベクトル $\boldsymbol{\beta}_{\mathrm{A}}$ の最小二乗推定量 $\hat{\boldsymbol{\beta}}_{\mathrm{A}}$ は，

$$\hat{\boldsymbol{\beta}}_{\mathrm{A}}=(\boldsymbol{X}_{\mathrm{A}}^T\boldsymbol{X}_{\mathrm{A}})^{-1}\boldsymbol{X}_{\mathrm{A}}^T\tilde{\boldsymbol{y}}_{\mathrm{A}}, \tag{2}$$

$$\begin{aligned}\tilde{\boldsymbol{y}}_{\mathrm{A}}&=(\tilde{y}_{\mathrm{A},1},\tilde{y}_{\mathrm{A},2},\cdots,\tilde{y}_{\mathrm{A},N})^T\\&=\boldsymbol{y}-\mathbf{1}\hat{\beta}_0-\boldsymbol{X}_{\mathrm{B}}\hat{\boldsymbol{\beta}}_{\mathrm{B}},\end{aligned} \tag{3}$$

で与えられる．式 (2)，(3) から $\hat{\boldsymbol{\beta}}_{\mathrm{A}}$ を推定するにあたり，Bob の推定パラメータデータ $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$ に加え，$\boldsymbol{X}_{\mathrm{B}}$ が必要になる．しかし，プライバシー保護の観点から他者の保持しているデータを直接用いることができない．そこで，Bob は所持するデータ $\boldsymbol{X}_{\mathrm{B}}$ を秘匿し，

$$\boldsymbol{\rho}_{\mathrm{B}}=(\rho_{\mathrm{B},1},\rho_{\mathrm{B},2},\cdots,\rho_{\mathrm{B},N})^T=\boldsymbol{X}_{\mathrm{B}}\hat{\boldsymbol{\beta}}_{\mathrm{B}}, \tag{4}$$

を計算して，Alice に送ることにする．このとき，Alice が所持してるデータ $\tilde{\boldsymbol{y}}_{\mathrm{A}}$ と $\boldsymbol{\rho}_{\mathrm{B}}$ を用いて Bob の所持するデータ $\boldsymbol{X}_{\mathrm{B}}$ を求めようとすると，$q\geq 2$ の場合，解が不定となり Bob のデータを一意に求めることができない．従って，式 (3) の計算時に式 (4) を用いて，

$$\tilde{\boldsymbol{y}}_{\mathrm{A}}=\boldsymbol{y}-\mathbf{1}\hat{\beta}_0-\boldsymbol{\rho}_{\mathrm{B}}, \tag{5}$$

とすることで，Bob のデータが秘匿された状態でも回帰係数が推定できる．また，Bob についても同様に $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$ を，

$$\hat{\boldsymbol{\beta}}_{\mathrm{B}}=(\boldsymbol{X}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}})^{-1}\boldsymbol{X}_{\mathrm{B}}^T\tilde{\boldsymbol{y}}_{\mathrm{B}}, \tag{6}$$

として算出できる．ただし，$\hat{\beta}_0$ は次式で与えられる．

$$\hat{\beta}_0=\frac{1}{N}\sum_{i=1}^{N}(y_i-\rho_{\mathrm{A},i}-\rho_{\mathrm{B},i}), \tag{7}$$

式 (2)-(7) で与えられる各回帰係数の推定には他者の推定結果を必要とする．そこで，須子らは推定結果を繰り返し送受信し，複数回に渡って分散計算を行うことで推定値を収束させる方法を示した．須子らによる回帰係数の推定プロトコルは以下で与えられる．

[回帰係数の推定プロトコル]

Step1) ランダムに初期値 $\hat{\beta}_0^{(0)},\hat{\boldsymbol{\beta}}_{\mathrm{B}}^{(0)}$ を設定し，$\hat{\beta}_0^{(0)},\boldsymbol{\rho}_{\mathrm{B}}^{(0)}$ を Alice に送る．$l=1$ とする．

Step2) Alice は，$\hat{\boldsymbol{\beta}}_{\mathrm{A}}^{(l)}$ を求め，$\boldsymbol{\rho}_{\mathrm{A}}^{(l)}$ を計算し Bob に送る．

Step3) Bob は，$\hat{\boldsymbol{\beta}}_{\mathrm{B}}^{(l)}$ を求め，$\hat{\beta}_0^{(l)}$ と $\boldsymbol{\rho}_{\mathrm{B}}^{(l)}$ を計算し Alice に送る．

Step4) 収束条件を満たさない場合は，$l=l+1$ として，**Step2** へ戻る．収束した場合は，$\hat{\boldsymbol{\beta}}^*=(\hat{\beta}_0^{(l)},\hat{\boldsymbol{\beta}}_{\mathrm{A}}^{(l)T},\hat{\boldsymbol{\beta}}_{\mathrm{B}}^{(l)T})^T$ を結果として出力する．　□

3　提案手法

須子らの手法では，全ての説明変数を用いた場合の回帰係数の推定方法のみが示されている．しかし，学習データが与えられたもとで未観測データを予測する問題を考えたとき，全ての説明変数を含んだ重回帰モデルが予測精度が一番高いモデルとは限らない．そこで，本研究では AIC 基準を用いた変数減少法による変数選択を行い，予測精度の高いモデルを構築するプロトコルを提案する．

いま，全ての変数を取り込んだフルモデルを M_0 として，説明変数を除去していく方法を考える．M_0 から任意の説明変数 k 個 $(0\leq k\leq p+q)$ を除去して得られるモデルの集合を $\mathcal{M}_k=\{m_1^k,m_2^k,\cdots,m_{{}_{p+q}\mathrm{C}_k}^k\}$ とする．モデル $m_v^k\in\mathcal{M}_k$ において Alice，Bob はそれぞれ $p_{m_v^k},q_{m_v^k}$ 個の説明変数を残しているとする．ここで，$p_{m_v^k}=1$，$q_{m_v^k}=1$ の場合，m_v^k に含まれる変数のデータが一意に求められるため，プライバシーが保護されない．そのため，$2\leq p_{m_v^k}\leq p$，$2\leq q_{m_v^k}\leq q$ とする．このとき，m_v^k で Alice と Bob がそれぞれ用いる説明変数データ行列は，$\boldsymbol{X}_{\mathrm{A},m_v^k}=[x_{i,s_{m_v^k}}]\in\mathbb{R}^{N\times p_{m_v^k}}$，$\boldsymbol{X}_{\mathrm{B},m_v^k}=[x_{i,t_{m_v^k}}]\in\mathbb{R}^{N\times q_{m_v^k}}$ となる．ただし，$s_{m_v^k}$ と $t_{m_v^k}$ は，m_v^k に含まれる説明変数の番号を指す．最尤推定した回帰係数ベク

トルを，$\hat{\boldsymbol{\beta}}^*_{m_v^k} = (\hat{\beta}^*_{0,m_v^k}, \hat{\boldsymbol{\beta}}^{*T}_{\mathrm{A},m_v^k}, \hat{\boldsymbol{\beta}}^{*T}_{\mathrm{B},m_v^k})^T \in \mathbb{R}^{1+p_{m_v^k}+q_{m_v^k}}$ とする．推定の際には，プライバシー保護のため以下の $\boldsymbol{\rho}_{\mathrm{A}m_v^k}, \boldsymbol{\rho}_{\mathrm{B}m_v^k}$ を計算し共有する．

$$\boldsymbol{\rho}_{\mathrm{A}m_v^k} = (\rho_{\mathrm{A}m_v^k,1}, \cdots, \rho_{\mathrm{A}m_v^k,N})^T = \boldsymbol{X}_{\mathrm{A},m_v^k} \hat{\boldsymbol{\beta}}_{\mathrm{A},m_v^k},$$
$$\boldsymbol{\rho}_{\mathrm{B}m_v^k} = (\rho_{\mathrm{B}m_v^k,1}, \cdots, \rho_{\mathrm{B}m_v^k,N})^T = \boldsymbol{X}_{\mathrm{B},m_v^k} \hat{\boldsymbol{\beta}}_{\mathrm{B},m_v^k}.$$

3.1 モデルの評価

実際に変数選択を行ったモデルに対して評価を行う必要がある．本研究では，その評価基準として未知データの予測を目的とした AIC を用いる．AIC は，式 (8) で与えられる．

$$\mathrm{AIC} = -2(\text{モデルの対数尤度}) + 2(\text{パラメータ数}). \quad (8)$$

モデル m_v^k において，推定した誤差分散を $\hat{\sigma}^2_{m_v^k}$ とすると，対数尤度 $l_{m_v^k}(\hat{\boldsymbol{\beta}}^*_{m_v^k}, \hat{\sigma}^2_{m_v^k})$ は以下で与えられる．

$$l_{m_v^k}(\hat{\boldsymbol{\beta}}^*_{m_v^k}, \hat{\sigma}^2_{m_v^k}) = -\frac{N}{2} \times \log(2\pi\hat{\sigma}^2_{m_v^k}) - \frac{N}{2}, \quad (9)$$

$$\hat{\sigma}^2_{m_v^k} = \frac{1}{N}\left\| \boldsymbol{y} - (\mathbf{1}\hat{\beta}^{(*)}_{0,m_v^k} + \boldsymbol{\rho}^*_{\mathrm{A}m_v^k} + \boldsymbol{\rho}^*_{\mathrm{B}m_v^k}) \right\|^2. \quad (10)$$

ただし，$\boldsymbol{\rho}^*_{\mathrm{A}m_v^k} = \boldsymbol{X}_{\mathrm{A}m_v^k} \hat{\boldsymbol{\beta}}^*_{\mathrm{A}m_v^k}, \boldsymbol{\rho}^*_{\mathrm{B}m_v^k} = \boldsymbol{X}_{\mathrm{B}m_v^k} \hat{\boldsymbol{\beta}}^*_{\mathrm{B}m_v^k}$ とする．このとき，モデル m_v^k のパラメータ数は定数項と誤差分散を含め $(p_{m_v^k}+q_{m_v^k}+2)$ 個であり，モデル m_v^k の $\mathrm{AIC}_{m_v^k}$ は，

$$\mathrm{AIC}_{m_v^k} = N \times \log(2\pi\hat{\sigma}^2_{m_v^k}) + N + 2(p_{m_v^k}+q_{m_v^k}+2), \quad (11)$$

となる．本研究では，AIC の値を小さくするように変数選択を行いモデルの構築を行う．

3.2 変数選択法

モデル m_v^k から $j_{m_v^k}$ 番目の変数を除去したときの AIC を AIC_{j,m_v^k} とする．Alice と Bob は，AIC_{j,m_v^k} が最小となる $\tilde{j}_{m_v^k}$ 番目の変数を除去する．これは，$\tilde{j}_{m_v^k}$ 番目の変数がモデル m_v^k に寄与していないと判断できるためである．ここで，式 (11) からパラメータ数が一定の時，AIC_{j,m_v^k} の値は誤差分散に寄与することに着目する．

そこで Alice は，モデル m_v^k において自分の所持する説明変数のうちモデル m_v^k に最も寄与していない $\tilde{s}_{m_v^k}$ 番目の変数を式 (12) を用いて求める．また，モデル m_v^k において $\tilde{s}_{m_v^k}$ 番目の変数を除いたときの誤差分散 $\sigma^2_{\tilde{s}_{m_v^k}}$ を式 (13) を用いて求める．

$$\tilde{s}_{m_v^k} = \underset{s_{m_v^k}}{\mathrm{argmin}} \sum_{i=1}^{N} (y_i - \rho_{\mathrm{A}m_v^k,i} - \rho_{\mathrm{B}m_v^k,i} + \hat{\beta}^*_{\mathrm{A},s_{m_v^k}} x_{i,s_{m_v^k}})^2, \quad (12)$$

$$\sigma^2_{\tilde{s}_{m_v^k}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \rho_{\mathrm{A}m_v^k,i} - \rho_{\mathrm{B}m_v^k,i} + \hat{\beta}^*_{\mathrm{A},\tilde{s}_{m_v^k}} x_{i,\tilde{s}_{m_v^k}})^2. \quad (13)$$

ただし，$\hat{\beta}^*_{\mathrm{A},s_{m_v^k}}$ は，$s_{m_v^k}$ に対応する回帰係数とする．また，Bob は $\tilde{t}_{m_v^k}$ 番目の変数と誤差分散 $\sigma^2_{\tilde{t}_{m_v^k}}$ を求める．Alice と Bob は，$\sigma^2_{\tilde{s}_{m_v^k}}, \sigma^2_{\tilde{t}_{m_v^k}}$ を比較する．$\sigma^2_{\tilde{s}_{m_v^k}} \leq \sigma^2_{\tilde{t}_{m_v^k}}$ ならば $\tilde{j}_{m_v^k} = \tilde{s}_{m_v^k}$ とし，$\tilde{s}_{m_v^k}$ 番目の変数を除去し，$p_{m_v^{k+1}} = p_{m_v^k} - 1, q_{m_v^{k+1}} = q_{m_v^k}$ とする．また，$\sigma^2_{\tilde{s}_{m_v^k}} > \sigma^2_{\tilde{t}_{m_v^k}}$ ならば $\tilde{j}_{m_v^k} = \tilde{t}_{m_v^k}$ とし，$\tilde{t}_{m_v^k}$ を除去し，$p_{m_{k+1_v}} = p_{m_v^k}, q_{m_v^{k+1}} = q_{m_v^k} - 1$ とする．ただし，須子ら [2] の手法と同様に，$\tilde{s}_{m_v^k}, \tilde{t}_{m_v^k}$ 番目の変数と誤差分散 $\sigma^2_{\tilde{s}_{m_v^k}}, \sigma^2_{\tilde{t}_{m_v^k}}$ の推定に直接互いのデータを用いずに $\boldsymbol{\rho}^*_{\mathrm{A}m_v^k}, \boldsymbol{\rho}^*_{\mathrm{B}m_v^k}$ を用いてプライバシーを保護する．

3.3 提案プロトコル

変数選択を行い，AIC が最小となるモデル m^* を選択するプロトコルを以下に示す．

[**提案プロトコル**]

Step1) $k = 0$ とする．

Step2) m_v^k について，$\hat{\boldsymbol{\beta}}^*_{m_v^k} = \left(\hat{\beta}^{(*)}_{0,m_v^k}, \hat{\boldsymbol{\beta}}^{(*)T}_{\mathrm{A},m_v^k}, \hat{\boldsymbol{\beta}}^{(*)T}_{\mathrm{B},m_v^k}\right)^T$ を推定し，$\mathrm{AIC}_{m_v^k}$ を求める．

Step3) $k > 1$ かつ，$\mathrm{AIC}_{m_v^k} > \mathrm{AIC}_{m_v^{k-1}}$ ならば，$m^* = m_v^{k-1}$ として終了．さもなくば，**Step4** へ．

Step4) m_v^k に最も寄与しない $\tilde{j}_{m_v^k}$ 番目の変数を選択する．

Step5) m_v^k から $\tilde{j}_{m_v^k}$ 番目の変数を除き，$p_{m_v^{k+1}} \geq 2, q_{m_v^{k+1}} \geq 2$ を満たせば $k = k + 1$ として，**Step2** へ．満たさなければ，$m^* = m_v^k$ として終了． □

Alice と Bob は，最終的にモデル m^* の回帰係数 $\hat{\boldsymbol{\beta}}^*_{m^*}$ を結果として共有する．

4 実験

4.1 実験条件

実験を行うにあたり，Alice と Bob が保持する説明変数をそれぞれ 20 変数とし，その中で，真のモデルに含まれる変数はそれぞれ 10 変数とする．説明変数と誤差は，正規乱数を用いて生成する．目的変数は，真のモデルに従い生成する．比較手法として，全ての説明変数を用いる須子ら [1] の手法を用いる．学習データ数は 100 件から 500 件の 100 件刻みとし，テストデータ 1,000 件に対して予測を行い，平均予測二乗誤差を計算する．これを 1 回として，1,000 回実験を繰り返し，その平均を結果として示す．

4.2 実験結果・考察

以下に実験結果を示す．

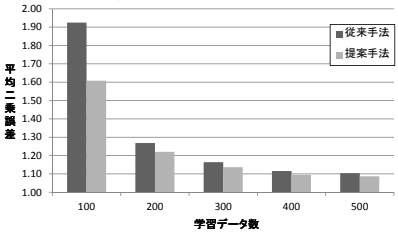

図 1：実験結果

図 1 から，全ての学習データで提案プロトコルが従来手法に比べて平均二乗誤差が小さいことが分かる．また，学習データが少ない場合は特に提案手法が有効であることが示された．

学習データ数が十分多い場合は，従来手法でも真のモデルに含まれない変数の偏回帰係数が 0 に近くなるため，提案手法との平均二乗誤差の差異が少なくなる．しかし，従来手法において学習データが少ない場合はオーバーフィッティングを起こしたため，変数選択を行う提案プロトコルとの平均二乗誤差の差異が大きくなったと考えられる．

5 まとめと今後の課題

本研究では，PPDM における線形回帰分析に対して，AIC を用いて変数選択を行うことで，予測精度の高いモデルを構築するプロトコルを提案し，有用性が示した．

今後の課題として，AIC が最小となるプライバシーを保護する変数選択法などが挙げられる．

参考文献

[1] 須子統太，堀井俊祐，小林学，後藤正幸，平澤茂一，“プライバシー保護を目的とした線形回帰モデルにおける最小二乗誤差の分散計算法について，” 信学技法，IBISML-2012