2002年日本オペレーションズ・リサーチ学会
秋季研究発表会

Support Vector Regressionによる顧客スコアリング

02103750　筑波大学　*後藤正輝　GOTO Masateru
01105930　筑波大学　香田正人　KODA Masato

1　はじめに

顧客スコアリングとは，過去の購買履歴データを基にして，購入可能性の高い順に顧客をランク付けする問題である．スコアリングモデルの学習は，ある期間に顧客が商品を購入するか否かという二値情報を，それ以前の購買行動から予測する回帰モデルを作成することにより行う．予測を行う顧客の購買履歴データをモデルに適用して得られた出力値を，購入可能性であるとみなし，その値を基にして顧客をランク付けする．

本研究ではSupport Vector Regression(以下SV回帰)を用いて，現実の企業の取引履歴データに対する顧客スコアリングモデルを作成した．その際スコアリング問題を，外れ値に対する頑健性が低いとされるＳＶ回帰に適合させるためのデータ前処理手法の提案を行う．

2　Support Vector Regression

学習しようとする入力$\mathbf{x} \in X \subseteq R^d$と出力$y \in R$の$l$個の組を訓練データSと呼び，$\mathbf{x}^{(i)}$を説明変数，$y^{(i)}$を教師信号と呼ぶ．

$$S = ((\mathbf{x}^{(1)}, y^{(1)}), \cdots, (\mathbf{x}^{(l)}, y^{(l)})) \in (X \times R)^l \quad (1)$$

高次元空間への非線型写像Φを考える．

$$\mathbf{x} = (x_1, \ldots, x_d) \mapsto \Phi(\mathbf{x}) = (\Phi_1(\mathbf{x}), \ldots, \Phi_N(\mathbf{x})) \quad (2)$$

$Im(\Phi)$を特徴空間と呼ぶ．SV回帰とは，特徴空間上で定義される線形モデル

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \Phi(\mathbf{x}) \rangle + b \quad (3)$$

によって，訓練データSを近似する問題である．損失関数としてϵ-insensitive損失関数

$$\begin{aligned} L^{\epsilon}(\mathbf{x}^{(i)}, y^{(i)}, f) &= |y^{(i)} - f(\mathbf{x}^{(i)})|_{\epsilon} \\ &= \max(0, |y^{(i)} - f(\mathbf{x}^{(i)})| - \epsilon) \end{aligned} \quad (4)$$

を使用する．SV回帰の主問題は以下のように表現される．

$$\begin{aligned} \min_{\mathbf{w}, \xi_i, \hat{\xi}_i} \quad & \| \mathbf{w} \|^2 + C \sum_{i=1}^{l} (\xi_i + \hat{\xi}_i) \\ \text{s.t.} \quad & \Big(\langle \mathbf{w} \cdot \Phi(\mathbf{x}^{(i)}) \rangle + b \Big) - y^{(i)} \leq \epsilon + \xi_i \\ & y^{(i)} - \Big(\langle \mathbf{w} \cdot \Phi(\mathbf{x}^{(i)}) \rangle + b \Big) \leq \epsilon + \hat{\xi}_i \\ & \xi_i, \hat{\xi}_i \geq 0, i = 1, \ldots, l \end{aligned} \quad (5)$$

ここで，ξ_iはマージンスラック変数$\xi_i = L^{\epsilon}(\mathbf{x}^{(i)}, y^{(i)}, f)$である．また，$\Phi$のカーネル関数を$K$とすると，SV回帰の双対問題は以下のように表すことができる．

$$\begin{aligned} \max_{\beta_{i,j}} \quad & \sum_{i=1}^{l} y^{(i)} \beta_i - \epsilon \sum_{i=1}^{l} |\beta_i| - \frac{1}{2} \sum_{i,j=1}^{l} \beta_i \beta_j K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \\ \text{s.t.} \quad & \sum_{i=1}^{l} \beta_i = 0 \ , \ -C \leq \beta_i \leq C \end{aligned} \quad (6)$$

3　データ前処理手法の提案

ＳＶ回帰は，モデルの出力値と教師信号の誤差をϵ以内に収めることを制約条件にもつ数理計画問題である．ϵより大きな誤差は目的関数にペナルティを与えた上で，制約条件がマージンスラック変数ξにより緩和される．しかしϵより大きな誤差を持つような学習データはすべてサポートベクターになり，その回帰モデルを構成する一要素になる．

スコアリング問題では両クラスの説明変数の分布が相互に重なり合っている．そのためＳＶ回帰によってスコアリングを行う場合，教師信号としてクラスラベル$\{1, -1\}$ = { 購入, 非購入 }をそのまま利用すると，多くの学習データがϵより大きな誤差をもつことになり，結果として必要以上に複雑なモデルが作られてしまう．

適切なモデルを作成するためには，近距離にあるデータが互いに近い教師信号の値を持つような学習データを構成する必要がある．また，クラス１に属するデータが多く分布する領域のデータは大きな教師信号の値をもち，クラス-1に属するデータが多く分布する領域のデータは小さな教師信号の値を持つ必要がある．

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

以上の条件を満たすような教師信号を作成する前処理手順を以下のように定義する．

1. すべての説明変数を $[-0.5, 0.5]$ の範囲に正規化する．

$$x_q^{(p)} = \frac{x_q^{(p)} - \min_i \arg x_q^{(i)}}{\max_i \arg x_q^{(i)} - \min_i \arg x_q^{(i)}} - 0.5 \qquad p = 1, \ldots, l \ , \ q = 1, \ldots, d \quad (7)$$

2. すべてのデータについて，Rより近い距離にある点の，クラスラベルの距離による加重和を求める．それをクラスラベルの絶対値の距離による加重和で割ることにより正規化する．ただし距離を $\frac{\|\mathbf{x}^{(p)} - \mathbf{x}^{(i)}\|}{\sqrt{d}}$ で定義することにより，正規化された説明変数についての最大距離を 1 に等しくする．

$$y^{(p)} = \frac{\sum_{i=1}^{l} y^{(i)} \max(0, 1 - \frac{\|\mathbf{x}^{(p)} - \mathbf{x}^{(i)}\|}{R\sqrt{d}})}{\sum_{i=1}^{l} \max(0, 1 - \frac{\|\mathbf{x}^{(p)} - \mathbf{x}^{(i)}\|}{R\sqrt{d}})} \qquad p = 1, \ldots, l \quad (8)$$

3. すべての教師信号の値を $[-1, 1]$ の範囲に正規化する．

$$y^{(p)} = 2\left(\frac{y^{(p)} - \min_i \arg y^{(i)}}{\max_i \arg y^{(i)} - \min_i \arg y^{(i)}}\right) - 1 \qquad p = 1, \ldots, l \quad (9)$$

4. 以上の手順で前処理を施した学習データでＳＶ回帰によるモデル作成を行う．

4 数値実験

説明変数が 2 次元である訓練データ [2] に対するスコアリングモデルの構築例を図 1 から 3 に示した．現実の企業の取引履歴データに対する顧客スコアリングモデルの構築については当日に報告する．

参考文献

[1] Alex J.Smola and Bernhard Schölkopf. A tutorial on support vector regression. Technical report, NeuroCOLT2 Technical Report Series, 1998.

[2] J.S. Simonoff. *Smoothing Methods in Statistic.* Spinger-Verlag Company, 1996.

[3] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[4] Stefan Rüping. *mySVM-Manual.* http://www-ai.cs.uni-dortmund.de /SOFTWARE /MYSVM /, 2000.

[5] 麻生英樹. 情報論的学習理論. **人工知能学会誌**, 16(2):287–297, 3 2001.

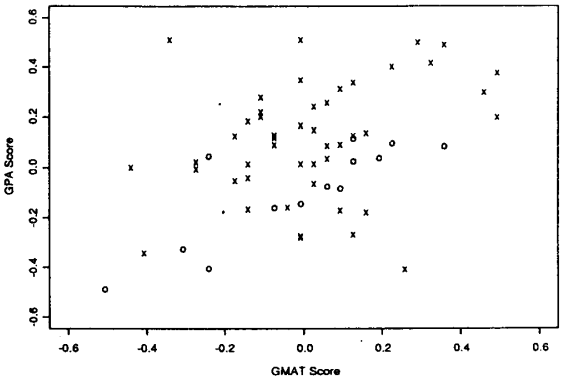

図 1: ＭＢＡ学生のＧＭＡＴ得点とＧＰＡのプロット．男性は×で表し，女性は○で表した．

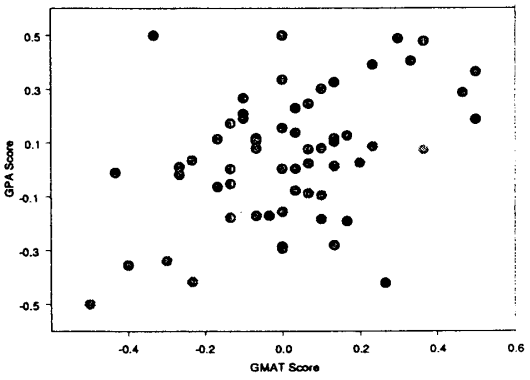

図 2: 前処理後のＭＢＡ学生のＧＭＡＴ得点とＧＰＡのプロット．平滑化パラメータは $R = 0.1$ である．淡色で示されているデータほど教師信号の値が小さく，女性である可能性が高いとみなされる．

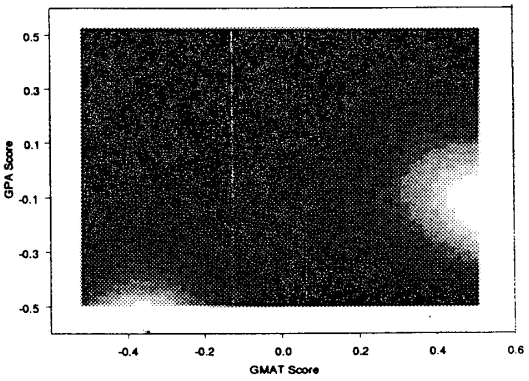

図 3: ＳＶ回帰による予測値．横軸がＧＭＡＴの得点で縦軸がＧＰＡの得点．学習パラメータは $C = 1000$, $\epsilon = 0.03$．またガウシアンカーネルの半径は $\gamma = 1.00$.

© 日本オペレーションズ・リサーチ学会. 無断複写・複製・転載を禁ず.