

2 値応答によるアンケート調査分析に対するアンサンブル学習法の開発

山梨大学	正会員	○下川 敏雄
山梨大学	正会員	大山 勲
山梨大学	正会員	風間ふたば
山梨大学	非会員	西山 志保
山梨大学	正会員	北村 眞一

1. 序に代えて

アンケートに代表される市民意識調査では、計量値でなく、計数値でとられることが多い。また、応答変数が2値でとられることも少なくない。このような場合には、重回帰分析ではなく、ロジスティック回帰分析を用いることが一般的である。ただし、ロジスティック回帰では、対数オッズに対して、説明変数が線形構造をもつため、真の構造が非線形の場合には適切でない。

下川他¹⁾は、アンケート調査結果に内在する交互作用効果を適切に捉えることができる回帰分析として、アンサンブル学習法の適用を推奨している。そこでは、とくに多重加法型回帰樹木法(MART)²⁾を取りあげ、MARTが既存の回帰分析諸法に比べて適合性能に優れ、そのグラフィカル診断法が結果に対する良好な示唆を与えることができることを指摘している。

ただし、既存のMARTは回帰モデルの枠組みで構成されている。そのため、本報告では、ロジスティック回帰の枠組みを用いることでMARTを2値応答データに拡張する。さらに、モデル・パラメータをLasso³⁾での縮小推定法を適用することで予測性能を向上させることができることを示す。このことは、MARTでのステージワイズ前進過程に対する後退あてはめ(モデル選択)に対応する。その結果は、長野県岡谷市および山梨大学の共同プロジェクトによって実施された水道水満足度アンケートの分析より得られた知見に基づいて評価する。

2. 2 値分類に対する縮小推定型アンサンブル学習法

MART法は、(CARTなどにより推定された)樹木モデルの線形結合(アンサンブル)によりモデルを構築する

回帰手法である。このとき、基本学習器パラメータは、 m 番目の樹木の「終結ふし」での部分集合 $t_q^{(m)}$ とその期待応答 $\gamma_q^{(m)} = E[\bar{y}(t_q^{(m)})]$ である($m=1, \dots, M$)。すなわち、MARTのモデルは、

$$F_{\text{boost}}^{(M)}(\mathbf{x}_i) = \sum_{m=1}^M \varsigma \sum_{q=1}^{Q_m} \gamma_q^{(m)} I(\mathbf{x}_i \in t_q^{(m)}) \quad (1)$$

により与えられる。ここに、 $\varsigma(0 \leq \varsigma \leq 1)$ は、縮小パラメータである。

MARTでは、ある微分可能な損失関数 $L(y, F_{\text{boost}}^{(M)})$ を定義し、それを最小化するように回帰パラメータ $\gamma_q^{(m)}$ を推定する。このとき、終結ふしでの部分集合 $t_q^{(m)}$ の推定には、損失関数の1次微分 $\partial L(y, F_{\text{boost}}^{(M)}) / \partial F_{\text{boost}}^{(M)}$ を用いる(疑似残差と呼ばれる)。

ただし、2値分類の場合には、通常回帰損失を用いることができないため、ロジスティック回帰に基づいてモデルを構成する。いま、説明変数 $\mathbf{x}_i$ および応答 $y_i \in (0, 1)$ が与えられたとき、群 $1(y_i = 1)$ に対する帰属確率は、 $p(\mathbf{x}_i) = \exp\{\mathbf{F}_{\text{boost}}^{(M)}(\mathbf{x}_i)\} / (1 + \exp\{\mathbf{F}_{\text{boost}}^{(M)}(\mathbf{x}_i)\})$ で与えられる。このとき、パラメータ $\gamma_q^{(m)}$ は、デヴィアンスに基づく損失関数の疑似残差 $r_{im} = y_i - p(\mathbf{x}_i)$ を用いることで

$$\gamma_q^{(m)} = \frac{\sum_{\mathbf{x}_i \in t_q^{(m)}} r_{im}}{2 \sum_{\mathbf{x}_i \in t_q^{(m)}} |r_{im}| (1 - |r_{im}|)} \quad (2)$$

により得ることができる。

MARTでは、ステージワイズ戦略に基づいてモデルが構築される。ただし、アンサンブルによって構成される樹木のなかで、不必要な樹木の影響を小さくし、重要な樹木の重みを強くすることは、適合性の向上お

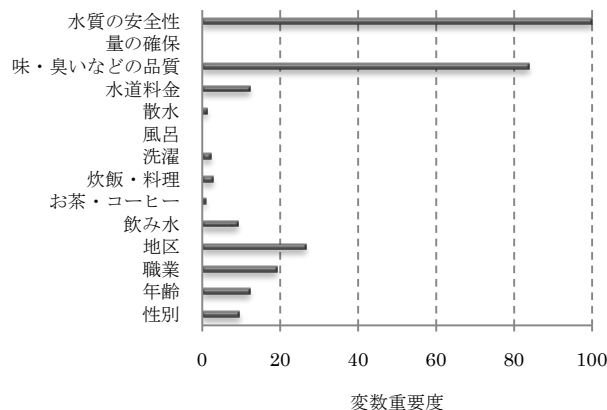


Fig. 1 L-MARTにおける変数重要度

よびモデルの安定化に繋がる。本報告では、lasso によるパラメータ推定を導入することでアンサンブルされた樹木への重みづけを行う。本報告では、提案手法を L-MART と呼ぶ。

3. 水道水満足度調査データの解析

長野県岡谷市住民の水道水に対する意識の解明のため、2006 年 1 月に、岡谷市と山梨大学が共同で郵送による質問紙調査を行った(抽出数 2000, 回答数 714, 回収率 35.7%)。そこでは、水道水に対する総合的満足度とともに、(a)水道料金、味・臭い、量の確保、安全性といった水道水に対する日常的印象、(b) 飲水、お茶・コーヒー、洗濯、風呂、散水といった用途、(c) 住民の年齢、性別、職業、あるいは居住地区といった回答者背景(属性)がとられた。

ここでは、水道水の総合的満足度に影響を与える要因を L-MART を通して解析する。Fig.1 は、L-MART における変数重要度(説明変数が応答に与える影響を表す指標)である。その結果、水質の安全性および品質が、水道水の総合的満足度に強い影響を及ぼしていることが示唆された。本アンケートでは、水質の安全性および品質に満足している回答者の 97.5%が水道水に満足している一方で、いずれにも満足していない回答者では、35.9%しか満足していなかった。また、地域による差異の変数重要度が 19%であり、3 番目に高かった。岡谷市では、高度処理水を主に利用している地域とそうでない地域が存在するが、高度処理水を利用している地域での水道水満足度が 81%であったのに対して、非利用地域では、91%であった(p 値=0.014)。水道水に対する安全性あるいは品質に関する重要性は、中西⁴⁾をはじめとした多くの文献・報告書で指摘されている。一方で、本解析では、岡谷市における取水地域による特

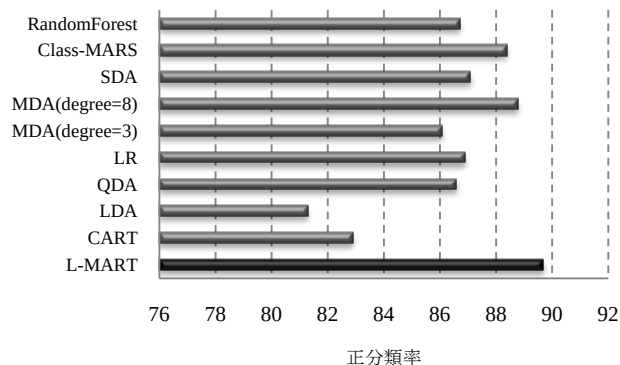


Fig. 2 他の手法との比較

(Class-MARS: 分類型多変量適応型回帰スプライン, SDA: 縮小推定型線形判別, MDA: 多変量正規混合判別, LR: ロジスティック回帰, QDA: 2 次判別, LDA: 線形判別)

性を明らかにしている。すなわち、L-MART の適用は、水道水満足度に対する、より詳細な要因構造を明らかにできる。

4. 他の手法との比較

Fig. 2 は、L-MART と他の手法での正分類率の比較である。線形モデルである。LDA, QDA, LR での正分類率が他の非線形モデルに比して低かった。また、LDA のパラメータ推定に縮小推定を用いた SDA では、正分類率が上昇した。さらに、CART での正分類率が最悪だったものの、アンサンブルすることにより、L-MART では、最良の正分類率を示している。

5. 結び

従来の調査研究、各都道府県市町村の水道事業評価意識調査は多く、調査項目も本調査と類似している、しかしその要因分析は一般に単純集計比較、重相関分析、分散分析であり、総じて線形モデルによる解析のみに終始している。

本研究では、L-MART の導入により、①手法としての新規性があり、②2 値への手法開発を行い、③河川満足度の地域特性などの新しい知見を得ており、④以上から手法の適用の有効性を示すことができた。

参考文献

- 1) 下川 敏雄, 武藤由香里, 御園生拓, 北村 眞一: 土木学会 景観・デザイン研究論文集, No.6, 2009(掲載予定).
- 2) Friedman, J.H: *Annals of Statistics*, Vol.29, pp.1189-1232, 2001.
- 3) Tibshirani, R. J. : *Royal. Statist. Soc. B.* , Vol.58, pp.267-288, 1996.
- 4) 中西 弘: 土木学会論文集, Vol.692, No. 21, p.p. 1-11, 2001.