ルール・アンサンブル法の開発

下川敏雄 *

1　序に代えて

自動交互作用検出法 ([AID: Automatic Interaction Detection], Morgan & Sonquist, 1963) に端を発する樹木構造接近法は，情報技術の発展やニーズの拡大 (例えば，データ・マイニングやケモメトリックス) により，急速にその版図を広げている．このような流れのなかで，分類回帰樹木法 ([CART: Classification And Regression Tree], Breiman, *et al.*, 1984) あるいは多変量適応型回帰スプライン法 ([MARS: Multivariate Adaptive Regression Spline], Freidman, 1991) などの多くの手法が提案され，その効用を発揮している．

ただし，樹木構造接近法での予測確度が粗悪なことは，広く知られており，場合によっては線形回帰分析を下回ることもある．また，MARS 法は，CART 法よりも予測確度に優れているものの，単調変換に対して不変でなく，さらに，外れ値および多重共線性の影響を受けるため，CART 法ほどの繁栄にいたっていない．近年，機械学習の分野から，任意の弱い学習器 (例えば樹木モデル) を組み合わせることによって強力な予測性能を持つ学習器を構成する方法が開発されている．それらはアンサンブル学習法と呼ばれ，統計科学および機械学習の分野で活発に研究されている．現在，多くのアンサンブル学習法が提案されている．

一つは，樹木構造接近法のモデル構築におけるランダム性 (変数選択およびふし点選択) に注目する方法である．そこでは，複数のブートストラップ標本あるいは部分標本に樹木を当てはめ，多数の樹木を構成する．そして，任意の説明変数に対する予測 (推定) は，それぞれの期待値をとる．このような接近法をモデル平均化接近法 (MAA: Model Averaging Approach) と呼ぶ．MAA には，Bagging 法 (Breiman, 1996)，あるいは RandomForest(Breiman, 2001) などがある．もう一つは，ステイジワイズ過程により，ある単純な学習器の線形結合により強力な (予測性能の高い) 学習器を構成する方法である．この方法は，Boosting と呼ばれている．Boosting 樹木法の最初の提案は，Freund & Schapire(1996) による．その後，Friedman(2001, 2002) により，深く精査され，その拡張として，より高い性能をもつ Boosting 樹木法，すなわち多重加法型回帰樹木法 (Multiple Additive Regression Trees, Friedman, 2001) が提案されている．

また最近になって，上記のアンサンブル学習法が，基本学習器の線形結合であることに注目し，(1) 基本学習器の選択過程，および (2) 個々の基本学習器に対する回帰パラメータの推定過程，の 2 段階推定により構成する方法，すなわち ISLE 法 (Friedman & Popescu, 2003) が提案されている．このとき回帰パラメータの推定には，Lasso 法 (Tibshirani, 1995) を用いることで，不必要な樹木 (回帰パラメータが 0 の樹木) を削除することができる．いい換えれば，この 2 段階推定は，前進過程と後退過程をアンサンブル学習のなかにとり込むことを意味する．Friedman & Popescu(2003) は，ISLE 法を適用することで RandomForest 法および Bagging 法において大幅に性能を向上できることを指摘している．

アンサンブル学習法の多くが基本学習器に樹木構造接近法を利用している．その理由は，(1) 比較的簡単に基本学習器の構成が可能であることと，(2) 説明変数の応答に対する影響を変数重要度という要約統計量で解釈できること，(3) 計算速度が比較的高速なこと，にある (Friedman & Poescu, 2003)．ただし，アンサンブル学習法

*山梨大学 大学院医学工学総合研究部, e−mail:shimokawa@yamanashi.ac.jp

では，モデルが「ブラックボックス化」されるため，元来，樹木構造接近法がもつ解釈の平易さを犠牲にしている．Friedman & Popescu(2008) は，樹木をそのままアンサンブルさせるのではなく，樹木によって構成されるふし (ルール) をアンサンブルさせることで解釈が平易で，かつ予測力に優れた RuleFit 法を提案している．

ただし，RuleFit 法の十分な性能の評価，例えば，他のアンサンブル学習法（RandomForest 法や MART 法）との比較が十分に行われていない．そこで，本研究では，文献事例および数値検証を通して RuleFit 法の性能を評価し，その構成に影響を及ぼす要因を探索する．

2　ルール・アンサンブル法

アンサンブル学習法のモデルは，複数の単純な「弱い」基本学習器 (樹木) を何らかの形式で連結することで構成される．例えば，多重加法型回帰樹木法 (MART) では，予測値をステイジワイズ過程による基本学習器の線形結合により得られ，Bagging 法あるいは RandomForest 法では，個々の基本学習器の平均値により得られる．アンサンブル学習法は，単一の基本学習器に比べて，劇的にその性能を向上させるものの，モデルを「ブラックボックス化」するため，結果に対する解釈は困難である．Friedman & Popescu(2008) は，解釈可能性をもたせたアンサンブル型学習法として，RuleFit 法を提案している．RuleFit 法では，樹木あるいは単回帰をアンサンブルして得られる基本学習器を，樹木の場合には，個々のふし (終結ふしとは限らない) のプロフィール，そして単回帰の場合には，説明変数をモデルの要素として用いる．そして，個々の要素に対する回帰係数は，lasso 縮小推定により得られる．これにより，アンサンブル過程で得られた大量のルールによるモデルあてはめと不必要なルールの削除が行われる．さらに，RuleFit 法では，変数重要度は，CART 法で用いられているような代替変数に基づく方法ではなく，ルールに対する回帰係数に基づいて算出される．このことは，変数の重要度だけでなく，ルールに対する重要度も得ることにも繋がる．

2.1　概要

アンサンブル学習法のモデルは，基本学習器の線形結合

$$F(\boldsymbol{x}) = \alpha_0 + \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}), \tag{1}$$

で与えられる．ここに，M はアンサンブルの大きさであり，$f_m(\boldsymbol{x})$ は説明変数 $\boldsymbol{x}$ より得られる基本学習器 (樹木) である．回帰パラメータ $\{\alpha_m\}_{m=1}^{M}$ は，RandomForest 法あるいは Bagging 法では $\alpha_m = 1/M$ である．

Friedman & Popescu(2003) は，式 (1) が基本学習器 $\{f_m(\boldsymbol{x})\}_1^M$ が与えられたもとで，基本学習器に対する線形結合と見なすことができることに注目し，基本学習器の推定 (モデル構築過程) と回帰パラメータの推定過程の 2 段階推定の方法，すなわち ISLE 法を提案している．このとき，回帰パラメータ $\{\alpha_m\}_{m=1}^{M}$ は，lasso による縮小推定

$$\{\hat{\alpha}\}_0^M = \arg\min_{\{\alpha_m\}_0^M} \sum_{i=1}^{N} L\left(y_i, \alpha_0 + \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{x}_i)\right) + \lambda \cdot \sum_{m=1}^{M} |\alpha_m|, \tag{2}$$

により得られる．ここに，右辺の括弧内の第 1 項は，応答 y_i をモデル (1) で推定したときの損失関数である．これにより，応答 y の推定に対して不必要な基本学習器 $f_m(\boldsymbol{x})$ は削除 (回帰パラメータ $\alpha_m = 0$ になる) される．このことは，アンサンブル学習における基本学習器の「刈り込み」に繋がり，より安定したモデルの推定が期待できる．

RuleFit 法は，ISLE 法の 2 段階推定のアルゴリズムを採用し，基本学習器を樹木の代わりにふしにより得られるルール，あるいは線形項により構成される．このとき，モデル構築過程は，アルゴリズム 1 で与えられる．ここに，パラメータ $\boldsymbol{p}_m$ は，基本学習器 $f_m(\boldsymbol{x})$ を規定するためのパラメータである．すなわち，基本学習器が CART

アルゴリズム.1 ：アンサンブル生成アルゴリズム

1 $F_0(\boldsymbol{x}) = \arg\min_\alpha \sum_{i=1}^{N} L(y_i, \alpha)$

2 $m = 1$ から M まで続く　{

3 $\quad \boldsymbol{p}_m = \arg\min_{\boldsymbol{p}} \sum_{i \in R_m(\eta)} L(y_i, F_{m-1}(\boldsymbol{x}_i) + f(\boldsymbol{x}_i; \boldsymbol{p}))$

4 $\quad f_m(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{p}_m)$

5 $\quad F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \varsigma \cdot f_m(\boldsymbol{x})$

6 }

7 アンサンブル $= \{f_m(\boldsymbol{x})\}_1^M$

樹木の場合には，分岐変数と分岐点を表す．また，$R_m(\eta) \subset \{\boldsymbol{x}_i, y_i\}_{i=1}^N$ は，N 個の標本からの $\eta(\eta \leq N)$ 個の部分集合である．さらに，$\varsigma(0 \leq \varsigma \leq 1)$ は縮小パラメータである．アルゴリズム 1 は，既存のアンサンブル学習法を包括することができる．Bagging 法 (Breiman, 1996) は損失関数に平方誤差損失 $L(y, \hat{y}) = (y - \hat{y})^2$ 用い，$\varsigma = 0$ と設定される．そして，サイズ η の部分集合は，テスト標本あるいはブートストラップ法により与えられる．RandomForest(2001) は，Bagging 法の方法に，樹木選定におけるアンサンブル要素を付加することで (3 行目) 与えられる．MART 法 (Friedman, 2001) は任意の y に対する，諸種の損失関数 $L(y, \hat{y})$ を用いる．また，$\varsigma = 0.1$ および $\eta = N/2$ である．推定値は，M 個の樹木ブースティング $F_M(\boldsymbol{x})$ により与えられる．

RuleFit 法のアンサンブル過程では，MART 法 (Friedman, 2001) と同様に諸種の損失関数が用いられる．例えば，2 乗誤差損失

$$L(y, F) = (y - F)^2 \tag{3}$$

あるいは，Huber(1962) の損失

$$L(y, F) = \begin{cases} (y - F)^2/2, & |y - F| < \delta \\ \delta(|y - F| - \delta/2), & |y - F| \geq \delta \end{cases}, \tag{4}$$

である (Friedman & Popescu(2004)．ここに，δ は，絶対残差 $\{|y_i - F(\boldsymbol{x}_i)|\}_1^N$ の α 番目の分位点から得られる．

2.2 ルール・アンサンブル

RuleFit 法では，CART 法の分岐過程を用いることで，M 回のアンサンブルで $K(\geq M)$ 個のルールが構成される．いま，S_j を説明変数 x_j のとる全ての可能な値の集合 $(x_j \in S_j)$ とし，s_{jk} を S_j の部分集合とする $(s_{jk} \subseteq S_j)$．このとき，k 番目のルールは

$$r_k(\boldsymbol{x}) = \prod_{j=1}^{p} I(x_j \in s_{jk}), \tag{5}$$

で構成される．ここに，$I(\cdot)$ は，括弧内が真なら 1，偽なら 0 をとる．s_{jk} に対して，説明変数 x_j が順序尺度あるいは比尺度の場合には，部分集合は，区間 $s_{jk} = (x_{jk}^-, x_{jk}^+]$ により得られる．したがって，RuleFit 法における基本学習器のパラメータ $\boldsymbol{p}_k$ は，分岐変数および区間である．カテゴリカル変数 (名義尺度) の場合には，分岐されたカテゴリの部分集合である．

アルゴリズム 1 の基本学習器 $f(\boldsymbol{x}; \boldsymbol{p})$ の生成に CART 法の分岐過程を用いることで，樹木 $\{f_m(\boldsymbol{x})\}_{m=1}^M$ のそれぞれのふし (内部および終結ふし) により，形式 (5) のルールを得ることができる．

このとき，M 回のアンサンブルにより得られる，ルールの合計 K は

$$K = \sum_{m=1}^{M} 2(t_m - 1), \tag{6}$$

である．ここに，t_m は終結ふしの数である．式 (1) より，ルール項のみによる RuleFit 法の推定モデルは

$$\hat{F}_{\mathrm{RFit}}(\boldsymbol{x}) = \hat{\alpha}_0 + \sum_{k=1}^{K} \hat{\alpha}_k r_k(\boldsymbol{x}), \tag{7}$$

である．また，パラメータ推定値 $\hat{a}_k$ は，式 (2) より

$$\{\hat{\alpha}_k\}_0^K = \arg\min_{\{\alpha_k\}_0^K} \sum_{i=1}^{N} L\left(y_i, \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(\boldsymbol{x}_k)\right) + \lambda \cdot \sum_{k=1}^{K} |\alpha_k|, \tag{8}$$

で与えられる．

このとき，式 (8) の第 2 項の lasso ペナルティは，ルールのサポートの影響を受ける．そのため，尺度依存性をもつため，その標準偏差より

$$sd_k = \sqrt{\varrho_k(1-\varrho_k)}, \tag{9}$$

によって規準化 $r_k \leftarrow r_k(\boldsymbol{x})/sd_k$ される．ここに，ϱ_k は学習標本におけるサポート

$$\varrho_k = \frac{1}{N}\sum_{i=1}^{N} r_k(\boldsymbol{x}_i), \tag{10}$$

である．

既存のアンサンブル学習法では，樹木 (基本学習器) の最大終結ふし数 (あるいは樹木の深さ) を予め設定する．RuleFit 法では，CART 法の分岐過程のみを用いる代わりに，終結ふし数を指数乱数に基づいて設定する．これにより，様々な大きさの樹木が構成できる．すなわち，m 番目のアンサンブルにより得られる樹木の終結ふし数 t_m は，

$$t_m = 2 + fl(\gamma),$$

で得られる．ここに，γ は，指数分布 $\mathrm{E}(1/(\bar{L}-2))$ に従う乱数であり，$fl(\gamma)$ は γ 以下の最大の整数である．さらに，$\bar{L}(\geq 2)$ は M 回のアンサンブルに対する終結ふしの期待値を表す．また，期待終結ふし数 $\bar{L}$ の選定には，交差確認法を用いることができる．

2.3 線形項の導入

既存のアンサンブル学習法では，単独の基本学習器に基づいているものの，複数の基本学習器を含むことに対する制約はない．樹木 (およびルール) による基本学習器は，潜在的に線形構造をもつ標的関数 (真のモデル) に対する適合は粗悪であり，複雑なルールを構成する．RuleFit 法では，個々の説明変数の線形関数を含めることを許容する．すなわち，RuleFit 法のモデル (7) が線形項を導入される．

ただし，RuleFit 法に線形項を含めることは，外れ値に比較的頑健な CART 法の特徴を阻害するおそれがある．そのため，Friedman & Popescu(2005) は，Huber の損失関数の類推より，修正型線形項

$$l_j(x_j) = \min(\delta_j^+, \max(\delta_j^-, x_j)), \tag{11}$$

を用いることを推奨している．ここに，δ_j^- および δ_j^+ は，外れ値とその他の観測値を区分するしきい値であり，変数 x_j の q および $(1-q)$ 分位点により得られる．このとき，Friedman & Popescu(2008) は $q \simeq 0.025$ を推奨している．

線形項を導入したときの，RuleFit 法による推定モデル (7) は

$$\hat{F}_{\mathrm{RFit}}(\boldsymbol{x}) = \hat{\alpha}_0 + \sum_{k=1}^{K} \hat{\alpha}_k r_k(\boldsymbol{x}) + \sum_{j=1}^{P} \hat{\beta}_j l_j(x_j) \tag{12}$$

で得られる．したがって，回帰パラメータ $\{\alpha_k\}_{k=0}^K$,$\{\beta_j\}_{j=1}^P$ は

$$(\{\hat{\alpha}_k\}_0^K, \{\hat{\beta}_j\}_1^P) = \mathrm{argmin}_{\{\alpha_k\}_0^K, \{\beta_j\}_{j=1}^P} \quad \sum_{i=1}^N \left(y_i, \alpha_0 + \sum_{k=1}^K \alpha_k r_k(\boldsymbol{x_i}) + \sum_{j=1}^n \beta_j l_j(x_{ij}) \right) + \lambda \cdot \left(\sum_{k=1}^K |\alpha_k| + \sum_{j=1}^P |\beta_j| \right), \tag{13}$$

により推定される．

ルール項と同様に，線形項の係数 β_j に対する lasso ペナルティも説明変数の尺度に依存するため，

$$l_j(x_j) \leftarrow 0.4 \cdot l_j(x_j)/\mathrm{std}(l_j(x_j)),$$

のように規準化され，式 (13) に用いられる．ここに，$\mathrm{std}(l_j(x_j))$ は学習標本における $l_j(x_j)$ の標準偏差であり，0.4 は，ルールのサポート (10) が一様分布 U(0,1) に従うと仮定したときの平均標準偏差 (9) である．

Friedman & Popescu(2008) は，数値検証により，線形項の有用性を指摘している．そのなかで，線形項の導入は，実際のデータ解析において，RuleFit 法の性能を飛躍的に向上させることがないものの，真のモデルの非線形性が強い場合にも，その性能を阻害することがないことが指摘されている．

3　グラフィカル診断

RuleFit 法の基本学習器は，樹木ではなく，ルール項あるいは線形項である．したがって，モデル全体の評価は困難なものの，モデルを構成するアンサンブル要素 (基本学習器) のそれぞれを評価することは平易である．このとき，有用な統計量がルール重要度である．ルール重要度は，個々のルールに対する回帰パラメータに基づいて得られる．言い換えれば，重回帰解析における偏回帰パラメータの類推と捉えることができる．また，RuleFit 法の変数重要度は，既存のアンサンブル型学習法あるいは CART 法と異なり，ルール重要度に基づいて構成される．さらに，任意の変数に対する他の変数との交互作用効果のグラフィカル診断にも応用できる．本節では，RuleFit 法における，諸種のグラフィカル診断の方法について触れる．

3.1　ルール重要度

回帰パラメータ α_k, β_j の推定過程において，予測器 (ルール項および線形項) の多くが 0 になることが知られている．Friedman & Popescu(2008) では，経験的に 80∼90%が削除されると述べている．他方，残存する基本学習器は，推定値に対して何らかの影響を与える．この影響力の強さは，回帰パラメータの推定値 $\hat{\alpha}_k$ および $\hat{\beta}_j$ で得られる．本報告ではこの測度をルール重要度と呼ぶ．ルール重要度は，推定回帰パラメータの絶対値に対して，その標準偏差を掛け合わせることにより得られる．したがって，ルール項 (5) の場合には

$$\mathscr{I}_k = |\hat{\alpha}_k| \cdot \sqrt{\varrho_k(1-\varrho_k)}, \tag{14}$$

で与えられる．ここに ϱ_k はルールのサポート (10) である．また，線形項 (11) の場合には，

$$\mathscr{I}_j = |\hat{\beta}_j| \cdot \mathrm{std}(l_j(x_j)), \tag{15}$$

である．ここに，$\mathrm{std}(l_j(x_j))$ は $l_j(x_j)$ の標準偏差である．通常の解釈では，ルール重要度は，式 (14) および (15) の最大値との割合（

さらに，ルール重要度は，任意の標本 $\boldsymbol{x}$ に対して算出することもできる．すなわち，ルール項の場合には

$$\mathscr{I}_k(\boldsymbol{x}) = |\hat{\alpha}_k| \cdot |r_k(\boldsymbol{x}) - \varrho_k|, \tag{16}$$

であり，また線形項の場合には

$$\mathscr{I}_j(x_j) = |\hat{\beta}_j| \cdot |l_j(x_j) - \bar{l}_j|, \tag{17}$$

である．ここに，$\bar{l}_j$ は $l_j(x_j)$ の平均値である．したがって，ルール重要度とは，対応する基本学習器 ($r_k(\boldsymbol{x})$ あるいは $l_j(x_j)$) が予測モデル (12) から削除され，切片 $\hat{\alpha}_0$ をルール項の場合に $\hat{\alpha}_0 \leftarrow \hat{\alpha}_k s_k$，線形項の場合に $\hat{\alpha}_0 \leftarrow \hat{\beta}_j \bar{l}_j$ のように調整したときの $F(\boldsymbol{x})$ の絶対変化量として捉えることができる．

上記の重要度を用いることで，説明変数空間の部分領域 S におけるルール重要度を定義することも可能である．すなわち，その重要度は，

$$\mathscr{I}_k(S) = \frac{1}{|S|} \sum_{x_i \in S} \mathscr{I}_k(\boldsymbol{x}_i), \quad \mathscr{I}_j(S) = \frac{1}{|S|} \sum_{x_i \in S} \mathscr{I}_j(x_{ij}), \tag{18}$$

である．ここに，$|S|$ は S に含まれる基本学習器の数である．通常は，説明変数よりも応答 (予測値) の部分領域に関心をもつことが多い．例えば，関心部分領域 S_y が境界 $F(\boldsymbol{x}_i) \geq \check{y}$ であるとき，部分領域 S_y は

$$S_y = \{\boldsymbol{x}_i | F(\boldsymbol{x}_i) \geq \check{y}\}, \tag{19}$$

である．

3.2 変数重要度

RuleFit 法では，他のアンサンブル学習法と同様に，モデル全体に対する説明変数の影響の評価はできない．このとき，Breiman *et al.*(1984) によって提案された，変数重要度 (および，最大の変数重要度との比率で表した相対的重要度) は，応答に対する個々の説明変数の影響を調査するのに有用である．

RuleFit 法における変数重要度は，CART 法での代理変数の概念に基づく変数重要度ではなく，基本学習器に基づいて定義される．したがって，説明変数 x_j を含む基本学習器のルール重要度が高いとき，その変数の変数重要度が高いと解釈される．

任意の標本 $\boldsymbol{x}_i$ での変数 x_{ji} における変数重要度 $\mathscr{V}_j(\boldsymbol{x}_i)$ は

$$\mathscr{V}_j(\boldsymbol{x}_i) = \mathscr{I}_j(x_{ji}) + \sum_{x_{ji} \in r_k} \mathscr{I}_k(\boldsymbol{x}_i)/p_k, \tag{20}$$

で与えられる．ここに，$\mathscr{I}_j(x_j)$ は，線形項 x_j での重要度であり，そして第 2 項は，x_j を含むルール $r_k(x_j \in r_k)$ の重要度 $\mathscr{I}_k(\boldsymbol{x})$ をそのルール内の説明変数の個数 p_k で割った値の総和である．

また，全標本 $\{\boldsymbol{x}_i\}_{i=1}^N$ における変数重要度 $\mathscr{V}_j$ は，式 (20) の全標本での平均値

$$\mathscr{V}_j = \frac{1}{N} \sum_{i=1}^{N} \mathscr{V}_j(\boldsymbol{x}_i)$$

により与えられる．同様に，部分領域 S での変数重要度 $\mathscr{V}_j(S)$ は

$$\mathscr{V}_j(S) = \frac{1}{N_S} \sum_{i \in S} \mathscr{V}_j(\boldsymbol{x}_i)$$

である．ここに，N_S は，部分領域 S に含まれる個体数である．

3.3 部分従属プロット

樹木あるいはルールに基づくアンサンブル学習法では，応答と説明変数の間の関数関係を平易に表すことができない．一方で，$\boldsymbol{x}$ と $\hat{y}$ を座標軸上にプロットし，視覚的に表現することは有用な示唆を与えることができる．$p > 3$ の場合には，グラフィカル表示ができないため，Friedman(2001) は，グラフィカル診断法として部分従属プロットを提案している．

いま，標本 $(y_i, \boldsymbol{x}_i), i = 1, \cdots, N$ (ここに，$\boldsymbol{x}_i = (x_{1i}, \cdots, x_{pi})^{\mathrm{T}}$ である) が与えられている．関心がある $p^+(< p)$ 個 (視覚的 に表現するためには，$p^+ = 1, 2$) の説明変数を $\boldsymbol{x}_s$ とし，それ以外を $\boldsymbol{x}_{\setminus s}$ とする ($\boldsymbol{x}_s \cup \boldsymbol{x}_{\setminus s} = \boldsymbol{x}$)．また，標本から得られる RuleFit 法の結果は $\hat{F}_{\mathrm{RFit}}(\boldsymbol{x}) = \hat{F}_{\mathrm{RFit}}(\boldsymbol{x}_s, \boldsymbol{x}_{\setminus s})$ である．このとき，$\boldsymbol{x}_{\setminus s}$ を固定したもとで $\hat{F}_{\mathrm{RFit}}(\boldsymbol{x}|\boldsymbol{x}_{\setminus s})$ が $\boldsymbol{x}_{\setminus s}$ のバラツキに強く影響を受けないという仮定のもとで，$\boldsymbol{x}_s$ に対する RuleFit 法の部分従属度 $PD_s(\boldsymbol{x}_s)$(Friedman, 2001) は

$$PD_s(\boldsymbol{x}_s) = \mathrm{E}_{\boldsymbol{x}_{\setminus s}}[\hat{F}_{\mathrm{RFit}}(\boldsymbol{x}_s, \boldsymbol{x}_{\setminus s}] = \int \hat{F}_{\mathrm{RFit}}(\boldsymbol{x}_s, \boldsymbol{x}_{\setminus s}) \pi_{\setminus s}(\boldsymbol{x}_{\setminus s}) d\boldsymbol{x}_{\setminus s} \tag{21}$$

と定義される．ここに，$\pi_{\setminus s}(\boldsymbol{x}_{\setminus s})$ は，$\boldsymbol{x}_{\setminus s}$ の周辺密度関数

$$\pi_{\setminus s}(\boldsymbol{x}_{\setminus s}) = \int \pi(\boldsymbol{x}) d\boldsymbol{x}_{\setminus s}$$

であり，$\pi(\boldsymbol{x})$ は $\boldsymbol{x}$ の同時密度関数である．実際には，データを $\hat{F}_{\mathrm{RFit}}(\boldsymbol{x})$ にあてはめ，部分従属度 (21) の経験推定値

$$\widehat{PD}_s(\boldsymbol{x}_s) = \frac{1}{N} \sum_{i=1}^{N} \hat{F}_{\mathrm{RFit}}(\boldsymbol{x}_s, \boldsymbol{x}_{\setminus s}$$

を用いる．このとき，部分従属プロットは，座標軸上に $(\boldsymbol{x}_{s,i}, \widehat{PD}_s(\boldsymbol{x}_{s,i})), i = 1, \cdots, N$ をプロットし，それぞれを折れ線グラフによって結びつけることで構成される．部分従属プロットは，暗箱となっている RuleFit 法での説明変数の応答への影響をグラフィカルに与えることができる．部分従属プロットの主たる関心は，$\boldsymbol{x}_s$ によって，応答がどのような影響を受けたかを視覚的に捉えることにある．そのため，Friedman(2001) では，$\widehat{PD}_s(\boldsymbol{x}_s)$ を中心化 ($\widehat{PD}_s(\boldsymbol{x}_s)$ の平均値を差分する) することを推奨している．

3.4 交互作用効果の摘出

Friedman & Popescu(2008) は，任意の説明変数 x_j とその他の変数のあいだの交互作用の大きさを表すための統計量（交互作用統計量）を提案している．この統計量は前節の部分従属度より得ることができる．いま，RuleFit 法のモデル $\hat{F}_{\mathrm{RFit}}$ が推定されたとき，2 個の説明変数 (x_j,x_k) に対する，2 次交互作用を診断するための統計量は，部分従属度の分散比

$$H_{jk}^2 = \frac{\sum_{i=1}^{N} \left[\widehat{PD}_{jk}(x_{ji}, x_{ki}) - \widehat{PD}_j(x_{ji}) - \widehat{PD}_k(x_{ki}) \right]^2}{\sum_{i=1}^{N} \widehat{PD}_{jk}^2(x_{ji}, x_{ki})} \tag{22}$$

により与えられる．もし，変数間に交互作用関係が存在しないとき，2 変数部分従属度 $\widehat{PD}(x_j, x_k)$ は個々の部分従属度の和 $\widehat{PD}(x_j, x_k) = \widehat{PD}(x_j) + \widehat{PD}(x_k)$ なので，式 (22) は 0 になる．さらに，x_j が他の変数とのあいだに何らかの 2 次交互作用効果をもつか否かを診断するための統計量は，

$$H_j^2 = \frac{\sum_{i=1}^{N} \left[\hat{F}_{\mathrm{RFit}}(\boldsymbol{x}_i) - \widehat{PD}_j(x_{ji}) - \widehat{PD}_{\setminus j}(x_{\setminus ji}) \right]^2}{\sum_{i=1}^{N} \hat{F}_{\mathrm{RFit}}(\boldsymbol{x}_i)} \tag{23}$$

である．ここに，$\widehat{PD}_{\setminus j}(x_{\setminus ji}), i = 1, 2, \ldots, N$ は，x_j 以外の変数における部分従属度である．

交互作用統計量 (22) および (23) に対する帰無分布は，帰無仮説 $\mathrm{H}_0 : F(\boldsymbol{x}) = F^{(1)}_{\mathrm{RFit}}(\boldsymbol{x})$ に基づいてパラメトリック・ブートストラップ法 (Efron & Tibshirani,1993) により推定される．ここに，$F(\boldsymbol{x})$ は真のモデル，そして $F^{(1)}_{\mathrm{RFit}}(\boldsymbol{x})$ は交互作用を含まない RuleFit モデルである．このとき，ブートストラップ標本は，交互作用を含まないモデル (帰無モデル) に対して，ランダムな残差を付与することにより

$$\tilde{y}_i = \hat{F}^{(1)}_{\mathrm{RFit}}(\boldsymbol{x}_i) + \left(y_{p(i)} - \hat{F}^{(1)}_{\mathrm{RFit}}(\boldsymbol{x}_{p(i)})\right)$$

で構成される．ここに，$\{p(i)\}_{i=1}^N$ は，整数 $\{1, 2, \ldots, N\}$ のランダムな順列である．したがって，帰無分布は，ブートストラップ標本 $\tilde{y}_i$ の経験分布 $\hat{g}(\tilde{y})$ により得られる．

4　事例検討および数値検証

ここでは，文献事例および数値検証を通して，RuleFit 法およびその診断方法の有用性について吟味する．

4.1　事例検討：ボストン住宅データ

Harrison & Rubinfeld(1978) は，大気汚染 (窒素酸化物濃度) が家の価格に影響を及ぼしているか否かに関して調査を行った．そこでは，ボストン市の調査地区 506 地区に関して，犯罪率 ($CRIM$), 25,000 平方フィート以上の宅地の保有率 (ZW), 小売店業以外の割合 ($INDUS$), チャールズ川沿いか否か ($CHAS$), 窒素酸化物の濃度 (NOX), 1 軒あたりの平均部屋数 (RM), 1940 年以前の建物の割合 (AGE), 中心街までの重みつき距離 (DIS), 幹線道路へのアクセスの良さ (RAI), 税率 (TAX), 生徒と教師の比率 (P/T), 黒人の割合 (B), 低所得者の割合 ($LSTAT$), 持ち家価格の中央値 (HM) が観測された．

本報告では，4 種類の基本学習器 (ルール項のみ，線形項のみ，加法モデル ($\bar{L} = 2$)，フルモデル ($\bar{L} = 4$)) を本データにあてはめた．最適なモデルを選定するために，平均絶対予測誤差

$$aae = \frac{\mathrm{E}_{\boldsymbol{x}y}|y - F(\boldsymbol{x})|}{\mathrm{E}_{\boldsymbol{x}y}|y - \mathrm{median}(y)|}$$

を 10 重交差確認法により推定した．その結果，ルール項のみの平均絶対予測誤差は $aae_{\mathrm{Rule}} = 2.036$ であり，線形項のみの平均絶対予測誤差は $aae_{\mathrm{Linear}} = 3.412$ であり，加法モデルでの平均絶対予測誤差は $aae_{\mathrm{Add}} = 2.450$ であり，そして，フルモデルでの平均絶対予測誤差は $aae_{\mathrm{Both}} = 2.031$ であった．したがって，本データでは交互作用効果を含む非線形モデルが適切なようである．Both において，ルール (あるいは線形) 相対的重要度が 15%以上のものを表 1 に示す．最も重要度が高かったのは，$LSTAT$ の線形関数 (低所得者の割合) であり，その影響は他のルール (あるいは線形) に比して顕著だった．回帰係数 $\hat{b}$ は，$LSTAT$ より，低所得者の割合が高くなるほど住宅価格が減少するようである．次いで，線形予測子 AGE(1940 年以前の建物の割合) の相対的重要度が高かった．

ルール項では，中心街までの重みつき距離 (DIS) が小さく，教員一人あたりの児童・生徒数の割合 P/T が高く，そして $LSTAT$ が非常に小さい地区での住宅価格の高さが示唆された．さらに，住宅の部屋数が多く (RM), 汚染が比較的進行していない (窒素酸化物濃度 NOX) 地区，および，住宅の部屋数が多く (RM)，P/T が低い地域でも，住宅価格が高くなる傾向にあった．一方で，住宅の部屋数が少なく，税率 (TAX) が高く，中心街から遠い地域は，P/T が高く，中心街に近くない地域とともに，より低価な家をもつ傾向にあった．

図 1 は，推定された RuleFit モデルでの個々の変数での相対重要度を表している．ルール重要度が極端に高かった $LSTAT$ の変数重要度が最も高く，次いで RM の影響度が強かった．2 番目に高いルール重要度をもつ AGE の変数重要度はそれほど高くかった．これは，非線形項 (ルール) に対する影響を持たないことが影響していると示唆される (このことは次節で述べる)．また，本解析の目標である NOX の影響の強さが示唆された．図 2 は，部

表 1: RuleFit 法により構成されたルール影響度の高い基本学習器

相対的重要度	係数	サポート	ルール
100	-0.40		線形：$LSTAT$
37	-0.036		線形：AGE
36	10.1	0.0099	$DIS < 1.40 \& P/T > 17.9 \& LSTAT < 10.5$
35	2.26	0.23	$RM > 6.62 \& NOX < 0.67$
26	-2.27	0.88	$RM < 7.45 \& DIS 1.37 \& TAX > 219.0$
25	-1.40	0.41	$DIS > 1.30 \& P/T > 19.4$
20	2.58	0.049	$RM > 7.44 \& P/T < 17.9$
19	1.30	0.21	$RM > 6.64 \& NOX < 0.67$
18	2.15	0.0057	$RM > 7.45 \& P/T < 19.7$

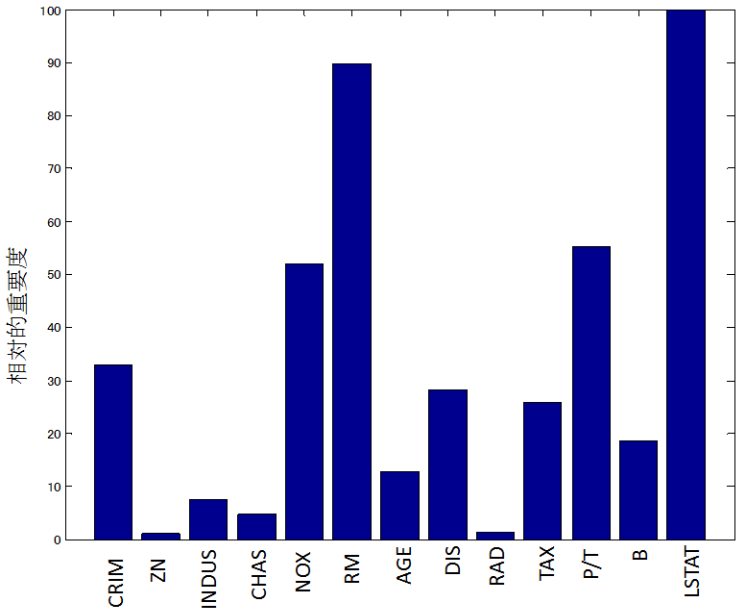


図 1: 相対的重要度の結果

分相対重要度を表している．左および中央のパネルは，住宅価格の予測値が，それぞれ下位 10%の地域，および中央値 ±5(%) の地域での変数重要度を表している．$LSTAT$ の影響度が全データのときに比べて極端に高いことから，中 ∼ 低所得者層の影響は，住宅価格が安い地域に対して強いようである．右のパネルは，住宅価格の上位 10%の地域での変数重要度を表している．ここでは，RM の影響が最も高く，P/T の影響も全データに比べて顕著に高かった．すなわち，住宅の大きさ，および教育環境が効果な住宅に強い影響を及ぼすようである．

次に，応答に対する個々の説明変数の影響をグラフィカルに省察した．図 3 は，1 変数部分従属プロットである．変数重要度の高い，RM および LSTAT の変動が大きかった．とくに，LSTAT は，おおよそ 5 以下で急激な上昇傾向が認められた．したがって，低所得者の割合が非常に少ない地域の住宅価格が急激に増加することが示唆される．また，RM では，6 部屋までの部分従属度は，ほぼ一定であるものの，7 部屋以上では急激な上昇傾向を示した．NOX は，0.6 付近にピークをもつ形状を示すものの，0.7 以上では非常に小さな値を示した．非常に公害が

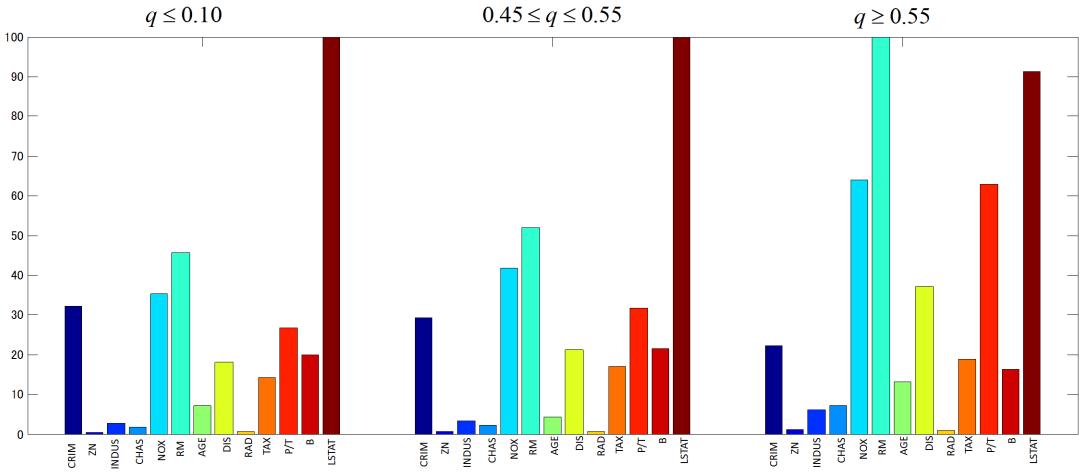


図 2: 部分相対的重要度の結果

進んだ地域での住宅価格は，低いようである．CRIM は，10 以上で急激な減少傾向を示した．いいかえれば，犯罪発生率の高い地域では，住宅価格が急激に減少する傾向にあった．また，中心地からの距離 (DIS) は，非常に近い地域では，大幅に住宅価格が上昇した．

平均絶対予測誤差において，本データの交互作用効果が示唆された．そのため，交互作用プロットにより，どの変数が交互作用をもつかを検討した．先ほどの 4 は，このときの交互作用プロットである．RM, P/T, LSTAT, NOX に高い交互作用効果が示唆された．そのため，これらの変数のすべての組み合わせに対して，2 変数部分従属プロットを描いた．NOX×RM での部分従属プロットでは，主効果の場合と同様に RM が 7 部屋以上で急激な上昇傾向がみられ，その傾向は，NOX が高いほど顕著だった (図 5(a))．P/T × NOX の部分従属プロットでは，強い交互作用関係が示唆されなかった (図 5(b))．RM×P/T の部分従属プロットでは，7 部屋以上で P/T が 14 以下で僅かな上昇傾向が認められた (図 5(c))．すなわち，規模の大きな住宅が多い地域では，生徒に対する教師の割合が高い地域ほど住宅価格が高いようである．LSTAT × NOX では，1 変量部分従属プロットでの LSTAT の急激な上昇と同様の傾向が認められるだけで，NOX の影響は殆どなかった (図 5(d))．RM$times$LSTAT では，部屋数が 7 部屋以上のときに，急激な上昇傾向が認められ，とくに LSTAT が小さいところではその傾向が顕著だった (図 5(e))．LSTAT と P/T では，LSTAT が小さいとき，P/T が 18 から 20 にかけて急激な減少傾向が認められた (図 5(f))．すなわち，低所得者の割合が低い地域では，生徒に対する教師の割合が低くなると住宅価格が減少する傾向にあることがわかった．

4.2 数値検証

目標とデザイン： 文献事例では，RuleFit 法とその診断手法の有用性を示した．RuleFit 法は，他の手法よりも少ないルールで良好なモデルを構築することができた．また，変数重要度および局所変数重要度は，応答に対する説明変数の影響のより詳細な吟味を支援することができた．さらに，交互作用プロットおよび部分従属プロットは，RuleFit 法のモデルに内在する説明変数間の交互作用を表すことができた．ただし，既存の樹木に基づくアンサンブル学習に比した性能評価には至っていない．

ここでは，(1) RuleFit 法に対する線形項の導入は適切か，(2) RuleFit 法は他の方法に比して予測性能 (確度) に優れているか，(3) 回帰パラメータに対する lasso 推定により，どの程度のルールが削除されたか，について評

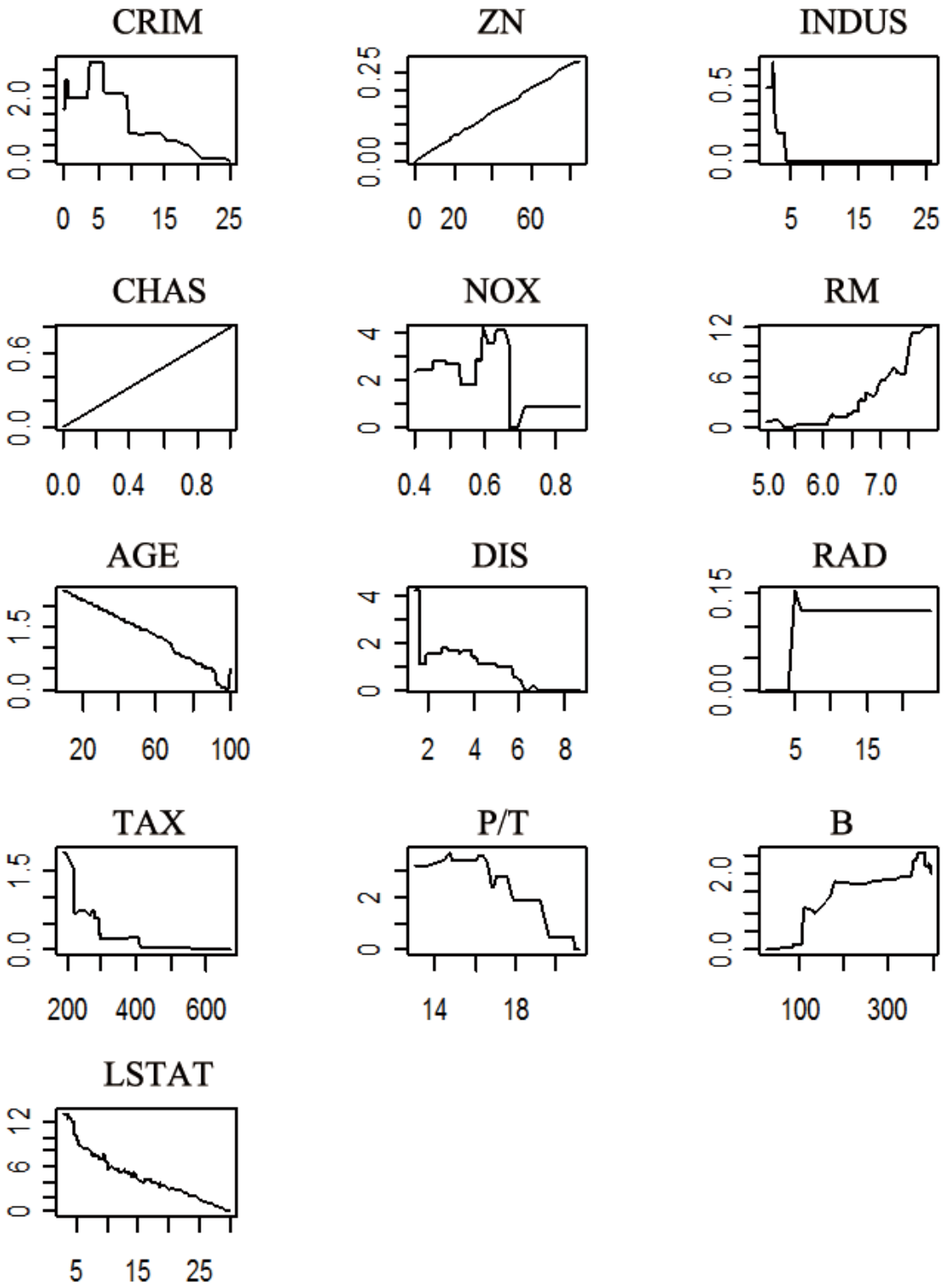


図 3: 1 変数部分従属プロット

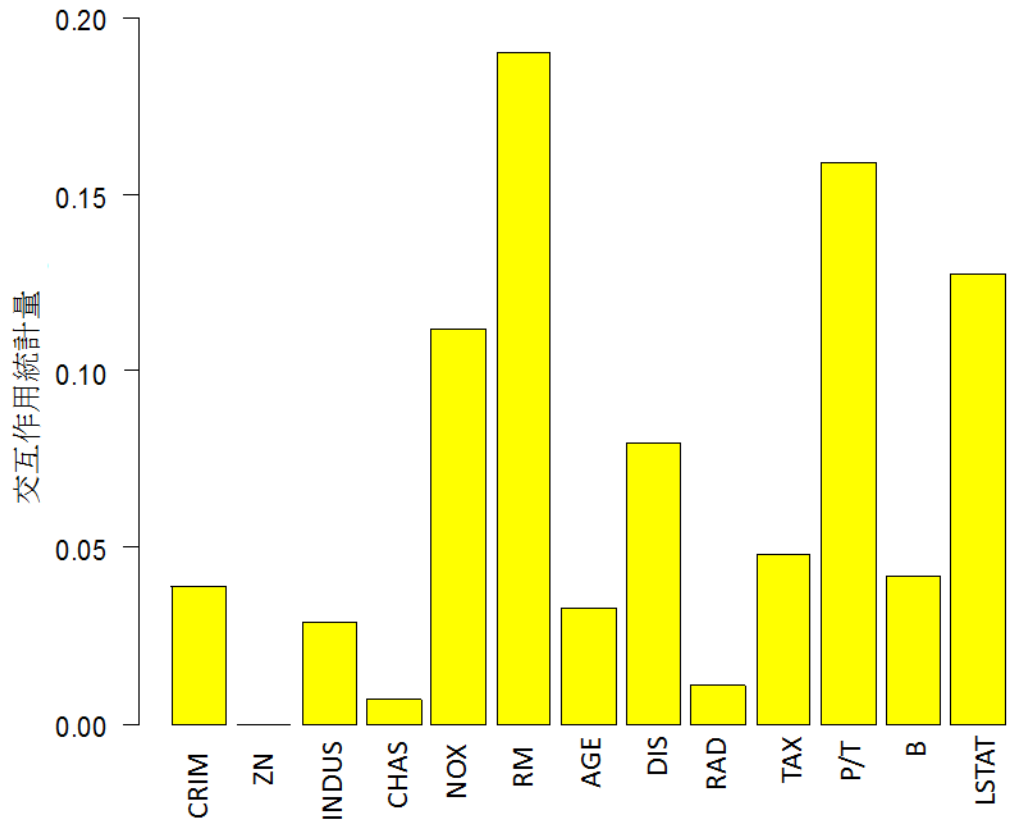


図 4: 交互作用プロット

価する．既存 (対照) の方法には，RandomForest 法，および MART 法を用いる．これらの方法を選択した理由は，アンサンブル学習のなかで最も有名であり，その有用性が広範に認められているためである．

真の (潜在的な) モデルとして，シミュレーション・モデルには

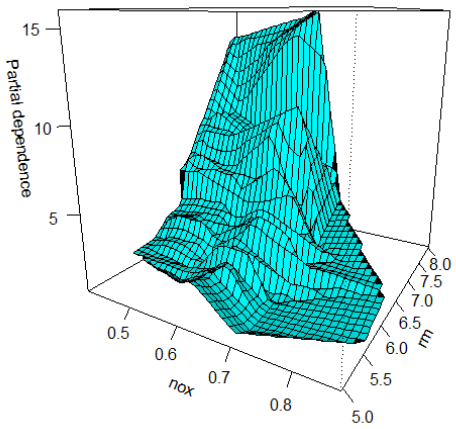
$$y = (1 - RL) \cdot (x_1 + x_2 + x_3 + x_4 + x_5) + \beta_0 \cdot RL \cdot (\iota_1 + \iota_2) + \epsilon \tag{24}$$

を用いる．ここに，ι_1，ι_2 はシグモイト関数

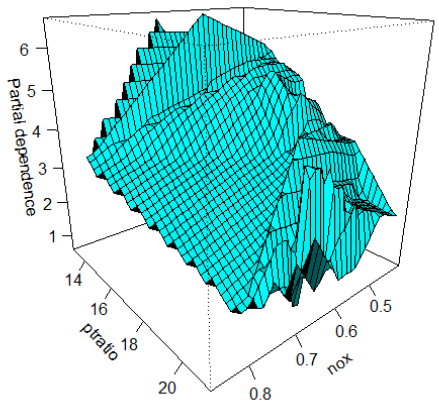
$$\iota_1 = \frac{1}{1 + \exp(x_1 + x_2 + x_3 + x_4 + x_5)}, \quad \iota_2 = \frac{1}{1 + \exp(x_6 - x_7 + x_8 - x_9 + x_{10})}$$

であり，誤差 ϵ は，標準正規分布 $\mathrm{N}(0, \sigma^2)$ に従う確率変数である．このとき，誤差分散 σ^2 は，全変動に占める回帰変動の割合 $RR(\%)$ により設定した．また，RL は，線形項と非線形項における変動の割合により設定した．すなわち，$RL = 0$ のとき，真のモデルは線形構造を示し，RL が増加するにつれて非線形構造の傾向が強くなり，$RL = 1$ のとき，真のモデルは完全な非線形構造を示す．β_0 は，$RL = 0.50$ のときに，線形項と非線形項の回帰変動が等しくなるように調整するためのパラメータである．

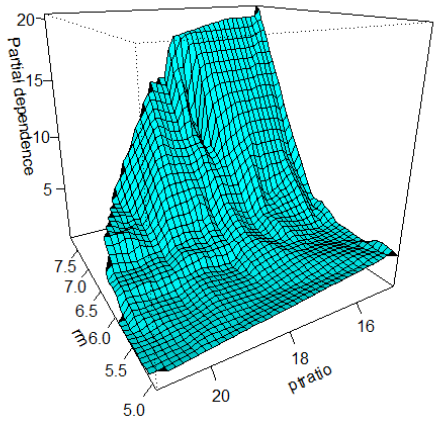
評価の方法： 本数値検証では，アンサンブル学習法に影響を及ぼす因子として標本サイズ N を N = 100,150, 200, 300, 500, 1000 の 6 水準，全変動に占める回帰変動の割合 RR を $RR = 0.7, 0.8, 0.9$ の 3 水準，そして線形/非線形の回帰変動の割合 RL を $0, 00, 0.50, 0.75, 1.00$ の 4 水準である．適用する方法として，RandomForest 法,



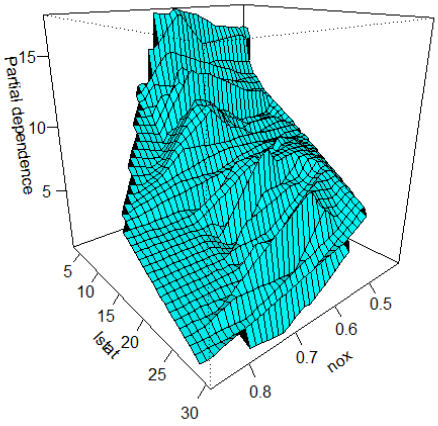
(a) NOX × RM



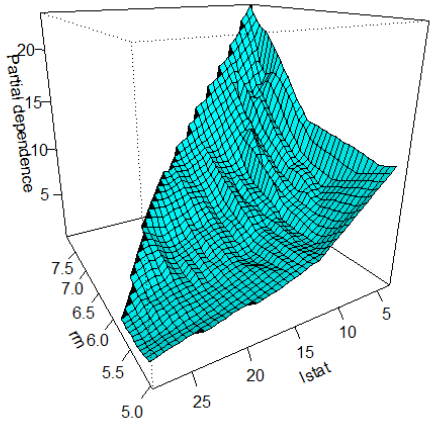
(b) NOX × PT



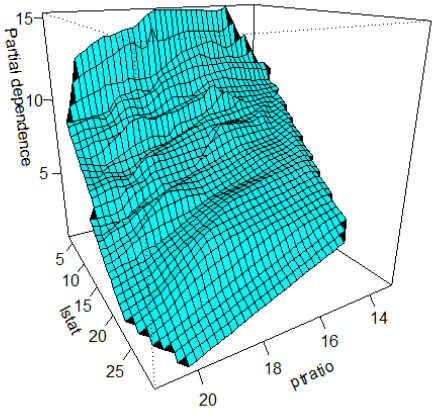
(c) P/T × RM



(d) NOX × LSTAT



(e) LSTAT × RM



(f) P/T × LSTAT

図 5: 2 変数部分従属プロット

MART 法, そして RuleFit 法を用いる．規定した因子のすべての組み合わせ ($6 \times 3 \times 4 = 72$ 通り) について，上記の設定のもとで学習標本を生成した．これらのデータに対して，RandomForest 法, MART 法, RuleFit 法を適用した．得られた結果の評価には，学習標本と同じ方法で生成された 500 個のテスト標本に対する平均平方誤差 MSE を用いた．そして，これを 500 回にわたって反復した．

結果： 図 6 は，数値検証の結果に対するウィンドゥ・プロットである．いずれの場合においても RandomForest 法の MSE は他の方法に比して極めて高かった．回帰問題において，RandomForest 法は良好な結果を示さないことが指摘されている (Breiman,2001; 杉本他,2005)．数値検証はそれを裏付ける結果を示した．真のモデルが完全な非線形構造をもつとき ($RL = 1.00$)，$N = 100$ において，RuleFit 法は MART 法よりも MSE が高かった．しかしながら，標本サイズが増加するにつれて MSE の差は縮まり，$N \geq 200$ において，RuleFit 法の MSE は MART 法を下回った．真のモデルに線形項が含まれるとき ($RL \leq 0.75$)，RuleFit 法は MART 法よりも良好な性能を示し，標本サイズが増加するにつれて，その差は広がる傾向にあった．標本サイズが大きい場合の 2 手法（MART 法，RuleFit 法）の差は，$RL = 0.50$ のときに最も顕著だった．$RL = 0.50$ の真のモデルは，線形項と非線形項が等しく影響を与える．いいかえれば最も複雑な構造をもつ場合である．したがって，RuleFit 法は真の構造が最も複雑な場合により効果を発揮するようである．

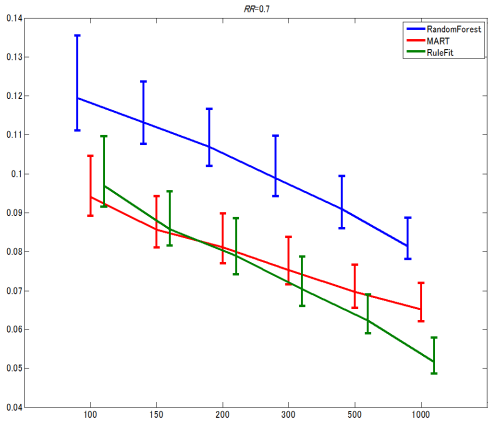
本数値検証の MSE は，真のモデルの尺度に依存するため，RuleFit 法に比したときの，他の手法の MSE を評価するほうが理解しやすい．表 2 は，数値検証の各サイクルにおける RuleFit 法の MSE に対する他の手法の MSE の割合を表している．したがって，1.00 よりも大きいとき，RuleFit 法のほうが良好な性能を示す．RandomForest 法は，いずれの場合にも RuleFit 法に比して粗悪だった．また，その傾向は，真のモデルが線形構造の影響が強くなるほど顕著だった．さらに，標本サイズが大きくなるにつれて，MSE の比が大きくなることが示唆された．真のモデルが完全な非線形構造をもち ($RL = 1.00$)，かつ $N = 100$ のとき，MART 法と RuleFit 法の MSE の割合が 1.00 を下回った．ただし，その他の場合には，1.00 を超えた．また，$RL = 0.50$ のときの MSE の比が最も高く，図 6 での示唆が裏付けられた．$N = 1000$ および $RL = 0.50$ のとき，RuleFit 法の MSE は，MART 法の約半分であり，RandomForest 法の約 1/3 であった．したがって，RuleFit 法は，他のアンサンブル学習法に比して極めて高い性能を示すことができた．

図 6 および表 2 では，$RL = 1.00$ かつ $N = 100$ のときに，MART 法よりも RuleFit 法のほうが MSE が高かった．そのため，$RL = 1.00$ での RuleFit 法で選択されたルールの割合を省察した (図 8)．その結果，アンサンブル過程で得られたルールの数パーセントしか推定モデルに含まれなかった．Friedman & Popescu(2008) では，10 ～ 20%のルールが含まれることを指摘していたが，本数値検証では，それを大きく下回った．これは，ここでの真のモデル構造が (非線形であるものの) それほど複雑でなく，かつ説明変数の数 p が $p = 10$ と比較的少なかったからであると推察される．標本サイズが少ない場合に，選択されたルールの割合のバラツキが大きく，標本サイズが上昇するにつれて，選択ルールの割合のバラツキが小さくなった．また，$N = 100$ から 200 のあいだで，選択されたルールの割合は緩やかな上昇傾向を示したが，$N = 300$ 以降で減少傾向を示した．$N = 200$ 以降では，標本サイズの設定幅を大きくしているため，厳密な傾向変化の位置がわからないものの，MSE の減少が顕著に認められているのが，$N \geq 300$ であることから，本モデルにおいて，RuleFit 法が予測確度に優れ，かつ安定した結果を示すには，$N \geq 300$ 以上が必要であることが示唆された．

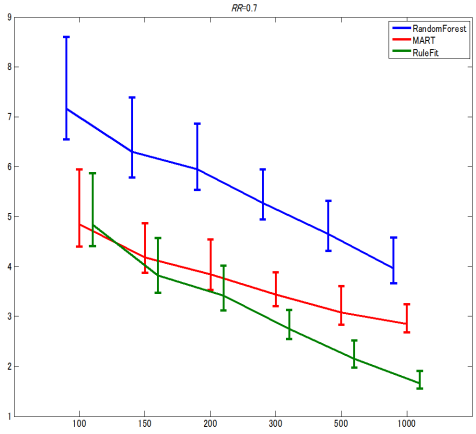
5　結びに代えて

本報告では，アンサンブル学習法の新たな接近法として，RuleFit 法およびそのグラフィカル診断法を俎上にあげ，その有用性について文献事例および数値検証により評価した．以下にその結果を要約する．

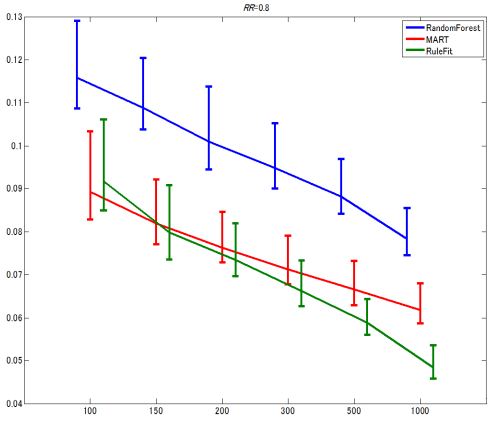
グラフィカル診断



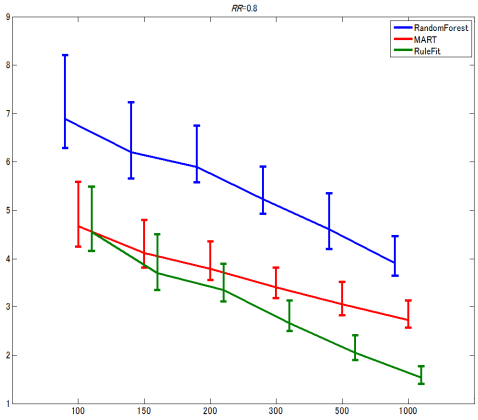
(a) $RL = 1.00, RR = 0.70$



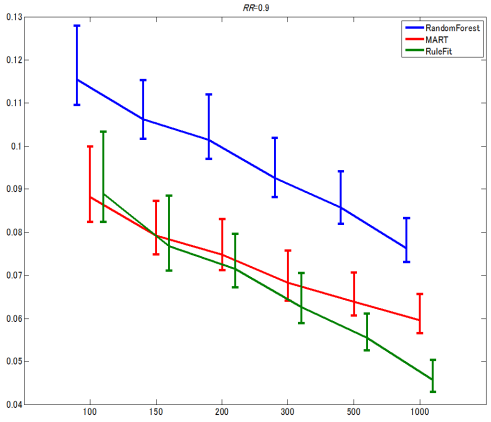
(b) $RL = 0.75, RR = 0.70$



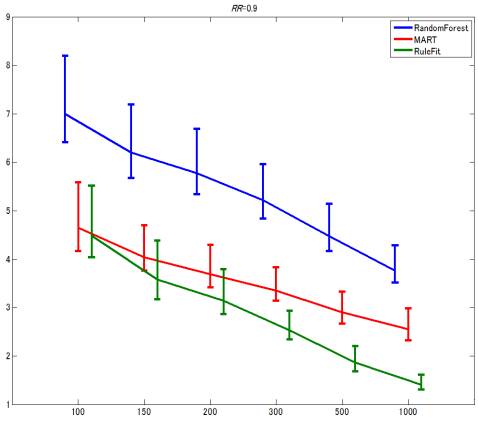
(c) $RL = 1.00, RR = 0.80$



(d) $RL = 0.75, RR = 0.80$

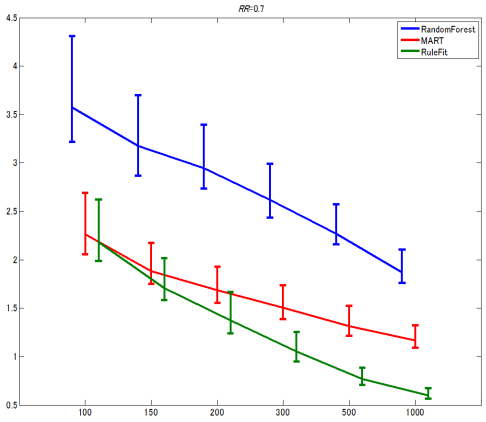


$RL = 1.00, RR = 0.90$

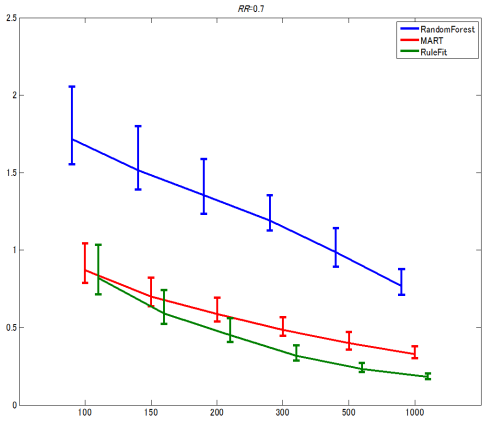


$RL = 0.75, RR = 0.90$

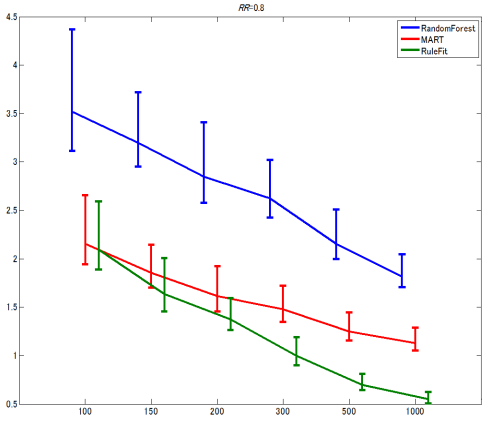
図 6: 数値検証の結果 (X 軸：標本サイズ，Y 軸：MSE)．ここにウィンドゥ幅は四分位範囲を表している．



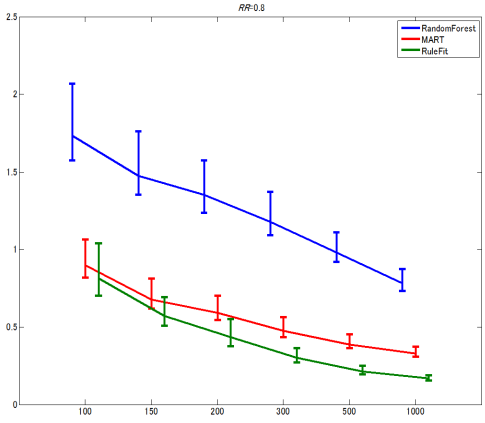
(e) $RL = 0.50, RR = 0.70$



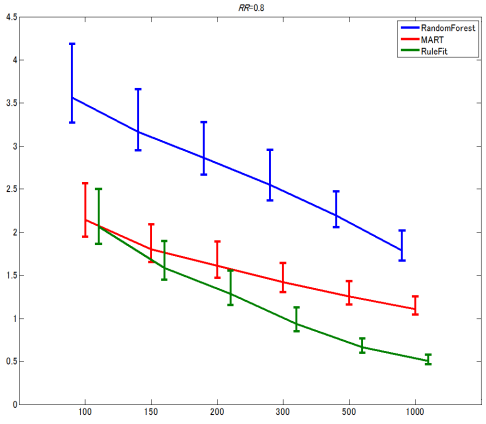
(f) $RL = 0.00, RR = 0.70$



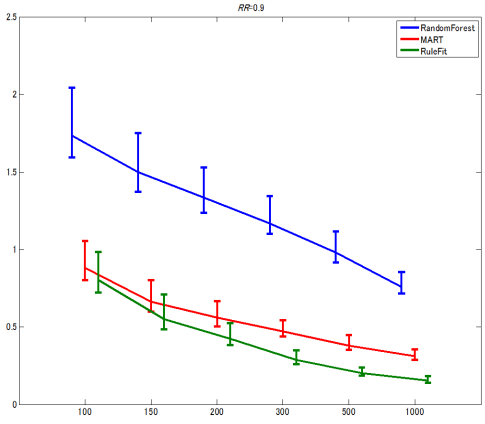
(g) $RL = 0.50, RR = 0.80$



(h) $RL = 0.00, RR = 0.80$



(i) $RL = 0.50, RR = 0.90$



(j) $RL = 0.00, RR = 0.90$

図 7: (続き) 数値検証の結果．ここにウィンドゥ幅は四分位範囲を表している．

表 2: RuleFit 法とその他の方法の MSE の割合 (X 軸：標本サイズ，Y 軸：MSE)

RL		1.00		0.75	
RR	N	RandomForest	MART		
0.7	100	1.23 [1.15 , 1.31]	0.97 [0.93 , 1.02]	1.48 [1.39 , 1.59]	1.00 [0.94 , 1.09]
	150	1.32 [1.25 , 1.37]	1.00 [0.97 , 1.04]	1.61 [1.51 , 1.75]	1.07 [1.01 , 1.13]
	200	1.34 [1.27 , 1.40]	1.02 [0.98 , 1.06]	1.72 [1.58 , 1.82]	1.11 [1.05 , 1.19]
	300	1.39 [1.32 , 1.47]	1.06 [1.02 , 1.10]	1.88 [1.78 , 2.04]	1.24 [1.18 , 1.32]
	500	1.45 [1.38 , 1.54]	1.11 [1.07 , 1.16]	2.14 [2.00 , 2.28]	1.44 [1.36 , 1.51]
	1000	1.55 [1.48 , 1.61]	1.24 [1.20 , 1.28]	2.38 [2.29 , 2.54]	1.71 [1.63 , 1.80]
0.8	100	1.25 [1.18 , 1.33]	0.98 [0.93 , 1.03]	1.49 [1.38 , 1.64]	1.03 [0.96 , 1.09]
	150	1.34 [1.27 , 1.40]	1.01 [0.97 , 1.06]	1.65 [1.55 , 1.75]	1.08 [1.02 , 1.15]
	200	1.38 [1.31 , 1.47]	1.04 [0.99 , 1.08]	1.72 [1.60 , 1.87]	1.13 [1.05 , 1.20]
	300	1.43 [1.36 , 1.51]	1.08 [1.04 , 1.12]	1.93 [1.82 , 2.08]	1.27 [1.19 , 1.32]
	500	1.50 [1.42 , 1.57]	1.13 [1.09 , 1.18]	2.24 [2.11 , 2.41]	1.50 [1.41 , 1.59]
	1000	1.60 [1.54 , 1.68]	1.28 [1.23 , 1.32]	2.51 [2.39 , 2.65]	1.76 [1.68 , 1.85]
0.9	100	1.26 [1.18 , 1.34]	0.98 [0.94 , 1.03]	1.53 [1.38 , 1.65]	1.03 [0.97 , 1.09]
	150	1.35 [1.27 , 1.43]	1.01 [0.97 , 1.07]	1.68 [1.57 , 1.84]	1.10 [1.03 , 1.16]
	200	1.41 [1.33 , 1.49]	1.04 [0.99 , 1.09]	1.84 [1.66 , 1.96]	1.16 [1.08 , 1.26]
	300	1.45 [1.39 , 1.52]	1.08 [1.04 , 1.12]	2.05 [1.90 , 2.18]	1.31 [1.24 , 1.38]
	500	1.54 [1.47 , 1.60]	1.16 [1.11 , 1.20]	2.35 [2.21 , 2.51]	1.54 [1.45 , 1.62]
	1000	1.67 [1.58 , 1.72]	1.31 [1.25 , 1.36]	2.68 [2.53 , 2.82]	1.85 [1.74 , 1.94]

RL		0.50		0.00	
RR	N	RandomForest	MART		
0.7	100	1.66 [1.50 , 1.81]	1.03 [0.96 , 1.11]	2.02 [1.82 , 2.25]	1.05 [0.95 , 1.12]
	150	1.86 [1.72 , 2.03]	1.09 [1.03 , 1.17]	2.49 [2.31 , 2.71]	1.15 [1.06 , 1.27]
	200	2.08 [1.88 , 2.25]	1.19 [1.11 , 1.28]	2.92 [2.68 , 3.21]	1.28 [1.16 , 1.38]
	300	2.45 [2.28 , 2.64]	1.42 [1.33 , 1.52]	3.65 [3.35 , 3.96]	1.51 [1.38 , 1.65]
	500	2.90 [2.71 , 3.14]	1.72 [1.61 , 1.82]	4.21 [3.93 , 4.50]	1.72 [1.62 , 1.84]
	1000	3.11 [2.96 , 3.29]	1.96 [1.85 , 2.06]	4.25 [3.99 , 4.45]	1.82 [1.73 , 1.92]
0.8	100	1.68 [1.54 , 1.85]	1.03 [0.96 , 1.11]	2.07 [1.87 , 2.28]	1.07 [0.97 , 1.20]
	150	1.88 [1.72 , 2.06]	1.10 [1.03 , 1.17]	2.57 [2.35 , 2.80]	1.18 [1.09 , 1.27]
	200	2.12 [1.95 , 2.32]	1.21 [1.12 , 1.31]	2.98 [2.71 , 3.28]	1.32 [1.19 , 1.42]
	300	2.64 [2.43 , 2.77]	1.48 [1.39 , 1.57]	3.88 [3.57 , 4.11]	1.57 [1.46 , 1.70]
	500	3.08 [2.86 , 3.28]	1.78 [1.69 , 1.88]	4.46 [4.16 , 4.88]	1.78 [1.67 , 1.92]
	1000	3.28 [3.09 , 3.47]	2.04 [1.94 , 2.17]	4.60 [4.29 , 4.94]	1.95 [1.81 , 2.08]
0.9	100	1.70 [1.55 , 1.85]	1.04 [0.96 , 1.11]	2.14 [1.89 , 2.36]	1.09 [1.00 , 1.17]
	150	1.96 [1.82 , 2.13]	1.13 [1.06 , 1.21]	2.59 [2.31 , 2.84]	1.18 [1.07 , 1.29]
	200	2.19 [2.01 , 2.34]	1.25 [1.15 , 1.32]	3.09 [2.81 , 3.37]	1.33 [1.19 , 1.44]
	300	2.65 [2.47 , 2.87]	1.49 [1.38 , 1.60]	3.94 [3.63 , 4.37]	1.60 [1.48 , 1.76]
	500	3.24 [3.03 , 3.48]	1.87 [1.76 , 1.98]	4.71 [4.34 , 5.19]	1.87 [1.73 , 2.04]
	1000	3.49 [3.29 , 3.71]	2.15 [2.04 , 2.28]	4.82 [4.46 , 5.24]	2.00 [1.87 , 2.15]

1. ルール重要度はアンサンブル学習に含まれる解釈の困難さを解消するのに有用だった.
2. 局所変数重要度は．関心のある応答の予測値に影響を与える説明変数の要因を明らかにできるため，PRIM 法 (Patient Rule Induction Method, Friedman & Fisher, 1999) のようなルール帰納法の代替あるいは包括手法として利用できることが示唆された.
3. 交互作用プロットは，任意の説明変数とその他のあいだの交互作用効果の強さをグラフィカルに提示できるだけでなく，部分従属プロットを構成するために有用な示唆を与えることがわかった.

数値検証

1. 真のモデルの非線形構造の傾向が強く，かつ標本サイズが少ない場合には，RuleFit 法に比して MART

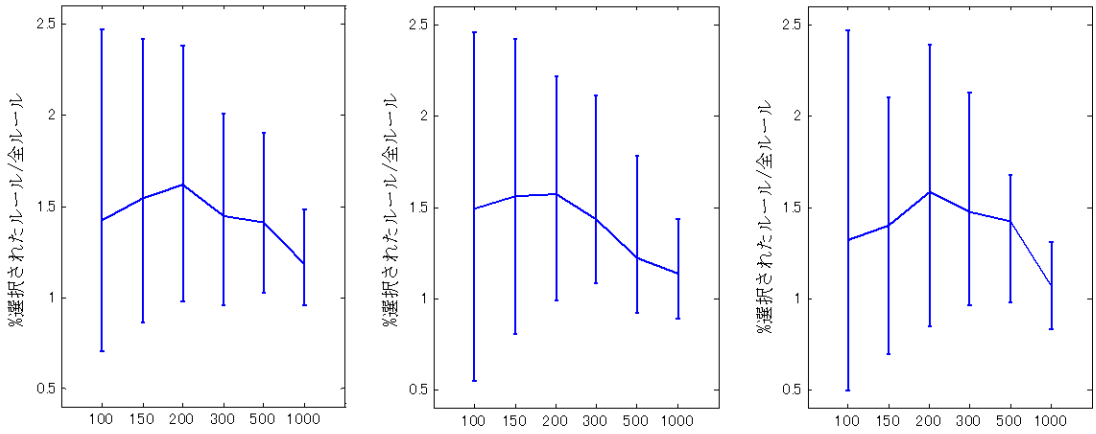


図 8: RuleFit 法における選択ルールの割合 (X 軸は標本サイズである)．ここにウィンドゥ幅は四分位範囲を表している．

法のほうが良好な性能を示した．ただし，標本サイズが増加するにつれて，RuleFit 法の MSE が MART 法を下回った．これは，RuleFit 法における lasso 法による回帰パラメータ選定過程が関連しているように思われる．RuleFit モデルのなかから削除されるルールは標本サイズが大きくなるにつれて多くなる傾向にある．すなわち，標本サイズの増加は，RuleFit 法により安定した予測モデルを提供する傾向にあり，このことがより良好なモデルを構成する結果に繋がったと推察される．

2. 真のモデルの線形傾向が強くなるほど，RuleFit 法が MART 法および RandomForest 法に比して良好な性能を示した．これは，RuleFit 法が線形項を許容しているためであると考えられ，シミュレーションは，線形項を導入することの適切性を裏付ける結果を示した．
3. いずれの場合においても，RandomForest 法が最も悪い性能を示した．杉本他 (2005) は，回帰問題において，RandomForest 法が良好な性能を示さないことを指摘しているが，シミュレーションは，このことを裏付ける結果だった．

RuleFit 法では，2 群分類に対する適用の示唆は与えられているものの，多群分類は考察されていない．他方，Hastie *et al.*(1994) は非線形多変量回帰モデルを多群判別分析に適用するための枠組み (柔軟判別解析法) を提案している．今後にかけて，RuleFit モデルを多応答に拡張したもとで多群判別に応用したい．さらに，RuleFit 法は既存のアンサンブル学習法に比して解釈が容易であることから，生存時間解析への応用が考えられる，Survival CART 法 (LeBranc & Crowley, 1992) を基本学習器とし，Survival lasso 法 (Tibshirani, 1997) を適用することで RuleFit 法が拡張できるか否かを検討したい．

参考文献

[1] Breiman, L., Friedman, J.H. Olshen, R.A. & Stone. C.J.(1984).*Classification and Regression Trees.* Wadsworth.

[2] Breiman, L. (1996). Bagging Predictors. *Machine Learning* , 26, 123–140.

[3] Breiman, L. (1998). Arcing classifiers (with discussion). *Ann. Statist.*, **26(3)**, 801–849.

[4] Breiman, L. (1999). Using adaptive bagging to debias regressions. Technical Report 547, Statistics Dept. UCB.

[5] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.

[6] Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap.* Chapman and Hall, London.

[7] Friedman, J. H (1991). Multivariate Adaptive Regression Splines (with discussion). *Annals of Statistics*, **19**, 1–141.

[8] Friedman, J. H., & Popescu, B. E. (2003). Importance Sampled Learning Ensembles.*Stanford University, Department of Statistics. Technical Report.*

[9] Friedman, J. H., & Popescu, B. E. (2004). Gradient directed regularization for linear regression and classification. *Stanford University, Department of Statistics. Technical report.*

[10] Friedman, J. H., & Popescu, B. E. (2008). Predivtive Learning via rule ensemble. *Ann. Appl. Stat.*, 2(3), 916–954.

[11] Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189 –1232.

[12] Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[13] Hastie, T., Tibshirani, R. & Buja, A. (1994). Flexible discriminant analysis by optimal scoring, *J. Amer. Statist. Assoc.* 89, 1255–1270.

[14] Hastie, T., Tibshirani, R. & Friedman. J.H.(2001). *The Elements of Statistical Learning: ata mining, inference and prediction.* Springer.

[15] Huber, P. (1964). Robust estimation of a location parameter. *Annals of Math. Statist.* 53, 73–101.

[16] LeBlanc, M. & Crowley, J.(1992). Relative risk trees for censored survival data. *Biometrics*, **48**, 411–425.

[17] Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, **5(2)**, 197–227

[18] Tibshirani, R. & Knight, K.(1999). Model search by bootstrap "Bumping", *Journal of Computational & Graphical Statistics*, 8 (4), 671–686.

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.* 58, 267–288.

[20] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model, *Statistics in Medicine*, 15, 385–395.

[21] Zhang, H. & Singer, B.(1999). *Recursive Partitioning in the Health Sciences.* Springer.

[22] 杉本知之・下川敏雄・後藤昌司 (2005). 樹木構造接近法と最近の発展. 計算機統計学, 計算機統計学, 18(2) ,123–164.