形項における変動の割合である．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．

このとき，説明変数 $x_1, \ldots, x_{15}$ は，一様乱数 $\mathrm{U}(0,1)$ により生成し，誤差分散 σ^2 は，全変動に占める回帰変動の割合 RR により設定する．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選定される．すなわち，N 個の乱数 $\boldsymbol{x}_i = (x_{1,i}, \ldots, x_{15,i})^{\mathrm{T}}, i = 1, 2, \ldots, N$ が生成されたもとで，式 (9) の右辺の第 1 項の平方和を V_L とし，第 2 項の平方和を V_{NL} とするとき，$\beta_0 = V_L/V_{NL}$ で与えられる．

本数値検証では，RuleFit 法および修正 RuleFit 法に影響を及ぼす因子として標本サイズ N を $N = 100, 200, 300, 400, 500$ の 5 水準，全変動に占める回帰変動の割合 RR を $RR = 6.0, 4.0, 2.0$ の 3 水準，そして線形/非線形の回帰変動の割合 RL を $0.25, 0.50, 0.75, 1.00$ の 4 水準を設定する．Elastic Net における L_1/L_2 割合 $\eta = 1.0, 0.8, 0.5, 0.2$ の 4 水準である．したがって，$\eta = 1.00$ のとき，通常の RuleFit 法と同様の結果が与えられる．規定した因子のすべての組み合わせ ($5 \times 4 \times 3 = 72$ 通り) について，上記の設定のもとで学習標本を生成し，修正 RuleFit 法を当てはめる．このとき，期待終結ふし数 $\bar{t}$ を $\bar{t} = 6$ (3 次交互作用) とし，縮小パラメータ λ は 10 重交差確認法により推定する．また，ルール項を構成するためのブースティング回数 M を $M = 500$ とする．

結果の評価には，Friedman and Popescu (2008) に倣い，学習標本と同じ方法で生成された 200 個のテスト標本に対する平均絶対偏差 ER_η

$$ER_\eta = \frac{1}{200} \sum_{i=1}^{200} |\hat{f}_\eta(\tilde{\boldsymbol{x}}_i) - \tilde{y}_i|$$

の最小値を分母としたときの割合

$$MAE_\eta = \frac{ER_\eta}{\min_\eta(ER_\eta)}, \quad \eta = 1.0, 0.8, 0.5, 0.2$$

および構成された基本学習器の数を用いる．ここに，$\{\tilde{\boldsymbol{x}}_i, \tilde{y}_i\}_{i=1}^{200}$ はテスト標本である．そして，シミュレート回数は，500 回とする．さらに，それぞれの因子が応答に及ぼす影響の大きさを 4 元分類分散分析により評価する．

表 1 は，数値検証における 4 元分類分散分析の結果である．標本サイズ (N) での寄与率が 27.4% であり最も高かった．次いで，Elastic Net における L_1/L_2 割合 η が 25.8% で高かった．すなわち，既存の RuleFit 法の lasso 法によるパラメータ推定を，より柔軟な Elastic Net 法に変更することは，予測確度に影響を与えることがわかった．線形/非線形の回帰変動の割合 (RL) での寄与率が 2.6% であることから，真のモデルの構造による強い影響が認められなかった．ただし，$RL \times \eta$ の 2 次交互作用での寄与率が 11.4% であることから，真のモデルの構造と Elastic Net における L_1/L_2 割合のあいだに交互作用が示唆された．

図 1(a) は，標本サイズ (N) と Elastic Net における L_1/L_2 割合 η の組み合わせ水準での MAE_η のウィンドゥ・プロット (ウィンドゥは，500 回の反復における MAE_η の平均値 $\pm$ 標準偏差を表している) である．いずれの場合でも，$\eta = 0.2$ のときに最も低かった．ただし，標本サイズが増加するにつれて，他のパラメータ ($\eta = 1.0, 0.8, 0.5$) での MAE_η が減少し，$N = 500$ では，ほぼ

表 1. MAE_η に対する 4 元分類分散分析の結果

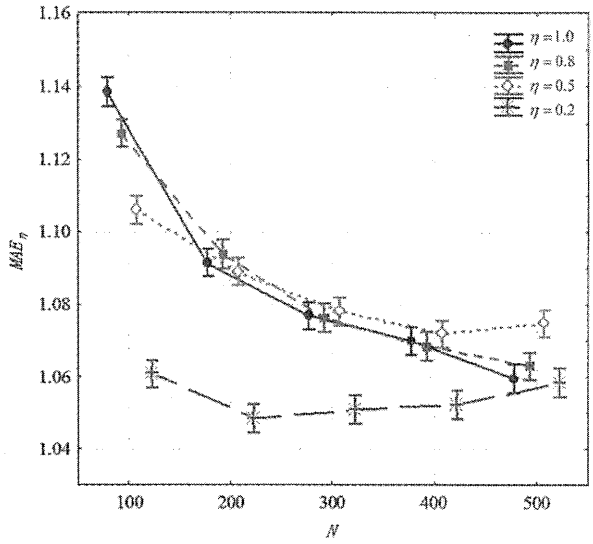
	p 値	寄与率
標本サイズ(N)	< 0.001	27.4
線形/非線形の回帰変動の割合(RL)	< 0.001	2.6
全変動に占める回帰変動の割合(RR)	< 0.001	18.2
Elastic Net における L_1/L_2 割合(η)	< 0.001	25.8
$RL \times \eta$	< 0.001	11.4
$N \times \eta$	< 0.001	3.6
$RR \times \eta$	< 0.001	3.9
$N \times RL$	0.827	0.1
$RL \times RR$	< 0.001	6.6
$N \times RR$	< 0.001	0.3

差異が認められなかった．通常 RuleFit 法($\eta = 1.0$)は，標本サイズが少ない場合の MAE_η が最も高かった．すなわち，標本サイズが少ない状況では，修正 RuleFit 法のほうが通常 RuleFit 法よりも良好な結果を示した．

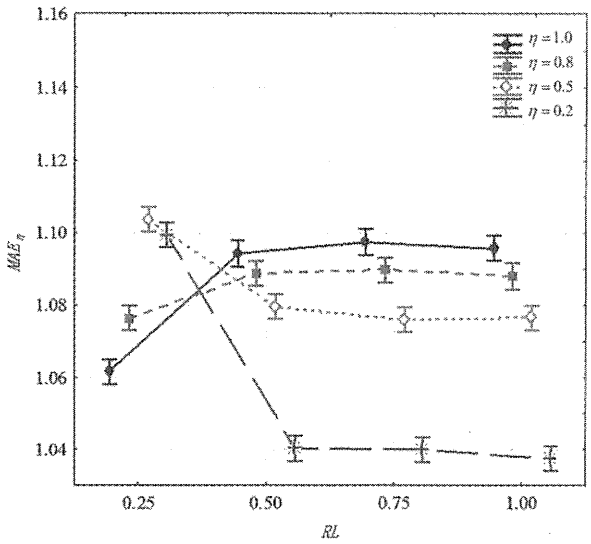
図 1(b) は，線形/非線形の回帰変動の割合(RL)と Elastic Net における L_1/L_2 割合 η の組み合わせ水準での MAE_η のウィンドゥ・プロット(ウィンドゥは，500 回の反復における MAE_η の平均値 ± 標準偏差を表している)である．真のモデルの線形傾向が強いとき($RL = 0.25$)，通常 RuleFit 法($\eta = 1.0$)での MAE_η が最も低かった．他方，$\eta = 0.2, 0.5$ といった，L_2 罰則の重みの強い修正 RuleFit 法の MAE_η は高かった．しかしながら，非線形傾向が強くなる(RL が高くなる)につれて，この傾向は逆転した．とくに，$\eta = 0.2$ での MAE_η は，$RL \geq 0.50$ において，極端に減少した．RuleFit 法では，他のアンサンブル学習法と異なり，基本学習器に線形項を含んでいることから，真のモデルの線形構造をもつ場合にも対応できる．$RL = 0.25$ のとき，通常 RuleFit 法($\eta = 1.0$)では，基本学習器を刈り込む傾向が強いため，影響力の弱い非線形項(ルール項)を修正 RuleFit 法よりも多く刈り込むことで，より良好なモデルを推定できたと推察できる．他方，真のモデルの非線形傾向が強くなる(RL が 1.00 に近づく)につれて，通常 RuleFit 法($\eta = 1.0$)では，過剰刈り込みに陥っていき，MAE_η が増加したと考えられる．他方，$\eta = 0.2$ では，過剰刈り込みの影響が少ないため，より良好な結果を示したと推察される．

図 1(c) は，全変動に占める回帰変動の割合(RR)と Elastic Net 罰則 η の組み合わせ水準での MAE_η のウィンドゥ・プロット(ウィンドゥは，500 回の反復における MAE_η の平均値 ± 標準偏差を表している)である．このプロットでは，誤差分散 σ^2 が大きくなるほど左方向になるように，X 軸の目盛り(RR の値)を逆に描写している．$RR = 6.0$ のとき(誤差分散が最も小さいとき)，すべての η が高い MAE_η を示し，かつ顕著な違いが認められた．これは，数値検証で生成されたデータの変動によって平均絶対偏差の最小値をもつ推定モデルが変化しているためである．いいかえれば，RR が高い状況では，安定して最も低い平均絶対偏差をもつ推定モデルが存在しないことを意味する．RR が減少するにつれて，$\eta = 0.2$ での MAE_η が減少した．すなわち，真のモデルのノイズが増大するほど修正 RuleFit 法の適切性が示唆された．前述したが，通常 RuleFit 法では，基本学習器を過剰に刈り込む傾向にある．数値検証の結果は，真のモデルに対するノイズが増加するにつれて，その傾向が顕著に現れることを示唆している．

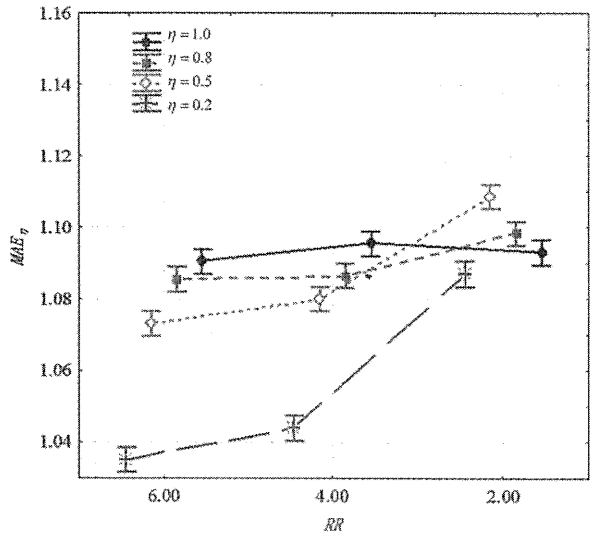
表 2 は，基本学習器の数に対する 4 元分類分散分析の結果である．MAE_η と同様に，標本サイ



(a) 標本サイズ(N)



(b) 線形/非線形の回帰変動の割合(RL)



(c) 全変動に占める回帰変動の割合(RR)

図 1. Elastic Net における L_1/L_2 割合 η と他の要因との組み合わせ水準における MAE_η のウィンドゥ・プロット

ズ(N)での寄与率が59.6%であり最も高かった．次いで，Elastic Net における L_1/L_2 割合 η が37.3%で高かった．すなわち，elastc net 罰則のパラメータ η によって基本学習器の数に差異が認められた．その他の要因および2次交互作用は，有意ではあるものの寄与率は非常に低かった．

MAE_η と同様に，η と他の要因との組み合わせ水準における2次交互作用のウィンドゥ・プロットを描いた(図 2)．いずれの場合にも，$\eta = 0.2$ での基本学習器の数が多く，その傾向は顕著だった．また，η の増加に伴い，基本学習器の数が減少する傾向にあった．

図 2(a) は，標本サイズ(N)と Elastic Net における L_1/L_2 割合 η の組み合わせ水準での基本学習器の数のウィンドゥ・プロットである．標本サイズが増加するにつれて，基本学習器の数が増加する傾向にあった．このとき，$\eta = 0.2$ の増加傾向は，他の修正 RuleFit 法に比べて僅かに急な勾配を示した．

図 2(b)(c) は，それぞれ，線形/非線形の回帰変動の割合(RL)と Elastic Net における L_1/L_2

表 2. 基本学習器の数に対する 4 元分類分散分析の結果

	p 値	寄与率
標本サイズ(N)	< 0.001	59.6
線形/非線形の回帰変動の割合(RL)	< 0.001	2.0
全変動に占める回帰変動の割合(RR)	< 0.001	0.1
Elastic Net における L_1/L_2 割合(η)	< 0.001	37.3
$RL \times \eta$	< 0.001	0.2
$N \times \eta$	< 0.001	0.0
$RR \times \eta$	< 0.001	0.1
$N \times RL$	< 0.001	0.4
$RL \times RR$	< 0.001	0.0
$N \times RR$	< 0.001	0.3

割合 η の組み合わせ水準，全変動に占める回帰変動の割合(RR)と Elastic Net 罰則 η の組み合わせ水準での基本学習器の数のウィンドゥ・プロットである．いずれのプロットにおいても，傾向変化が認められなかった．

数値検証の結果を以下に示す：(1) 標本サイズが少ない場合に修正 RuleFit 法は有用だった，(2) 真のモデルの線形傾向が強い場合には，通常 RuleFit 法は良好な性能を示したが，非線形傾向が増加するにつれて，通常 RuleFit 法では，過剰刈り込みの傾向を示した，(3) 真のモデルの誤差分散(ノイズ)が増加するにつれて修正 RuleFit 法の性能の適切性が示唆された，(4) η を減少させるにつれて，構成される基本学習器の数が増加傾向を示した，(5) 標本サイズが増加するにつれて，構成される基本学習器の数が増加傾向を示した．

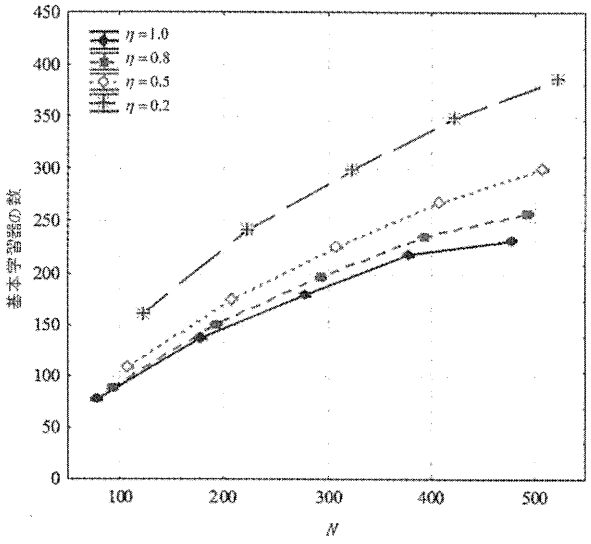
4.2. 他手法との比較

前節では，修正 RuleFit 法は，通常 RuleFit 法に比べて，良好なモデルを構成できることを数値検証により明らかにした．

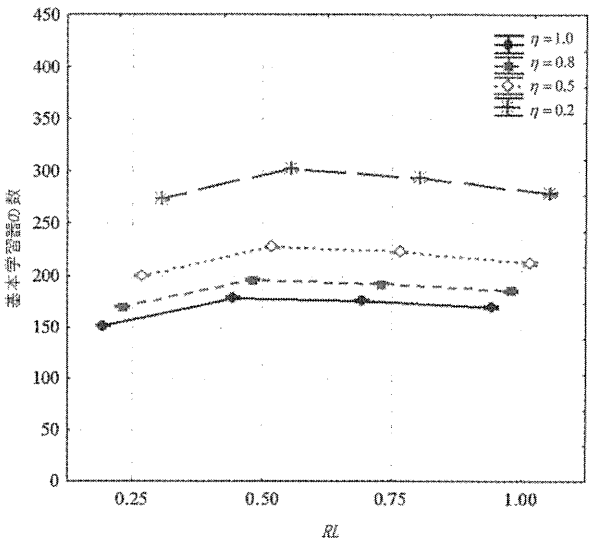
ここでは，修正 RuleFit 法は他の方法に比して予測性能(確度)に優れているか否かについて評価する．既存(対照)の方法には，RandomForest 法，および MART 法を用いる．これらの方法を選択した理由は，アンサンブル学習のなかで最も有名であり，その有用性が広範に認められているためである．

真の(潜在的な)モデルには前節の式 (9) を用いる．このとき，諸種の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 $N = 100, 200, 300, 400, 500$ の 5 水準，全変動に占める回帰変動の割合 RR を $RR = 6.0, 4.0, 2.0$ の 3 水準，そして線形/非線形の回帰変動の割合 RL を $0.25, 1.00$ の 2 水準に設定する．RL を 2 水準とした理由は，RandomForest 法および MART 法の基本学習器は樹木のみであるため，修正 RuleFit 法の優位性を $RL = 0.25$ において確認し，3 手法の性能比較を真のモデルに線形構造を含まない状況($RL = 1.00$)において評価するためである．

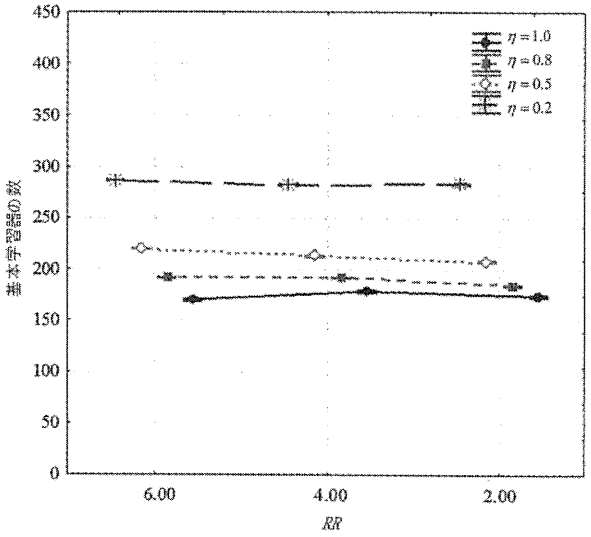
本数値検証でも，3.2 節と同様に，学習標本と同じ方法で生成された 200 個のテスト標本に対する平均絶対偏差を用いる．そして，修正 RuleFit 法での $ER_{\mathrm{EN-RuleFit}}$ に対する，RandomForest 法の平均絶対偏差 ER_{RF} および MART 法の平均絶対偏差 ER_{MART} の割合



(a) 標本サイズ(N)



(b) 線形/非線形の回帰変動の割合(RL)



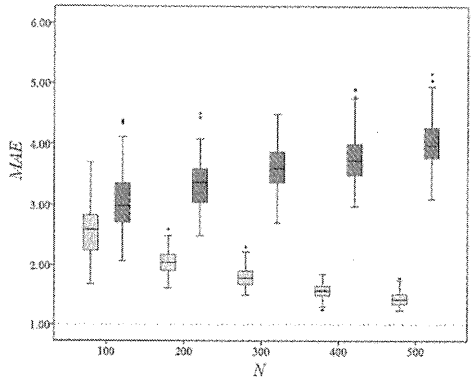
(c) 全変動に占める回帰変動の割合(RR)

図 2. **Elastic Net** における L_1/L_2 割合 η と他の要因との組み合わせ水準における基本学習器の数のウィンドウ・プロット

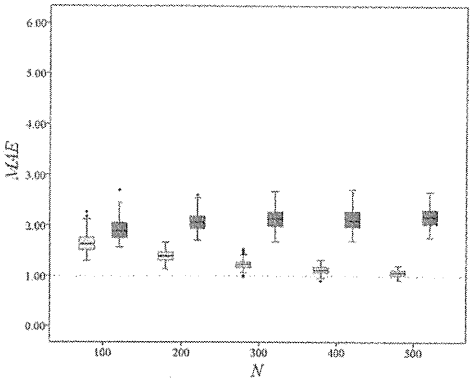
$$MAE_{\mathrm{RF}} = \frac{ER_{\mathrm{RF}}}{ER_{\mathrm{EN-RuleFit}}}, \quad MAE_{\mathrm{MART}} = \frac{ER_{\mathrm{MART}}}{ER_{\mathrm{EN-RuleFit}}}$$

により評価する．したがって，上式が 1.0 より大きいとき，既存のアンサンブル学習法は，修正 RuleFit 法よりも粗悪であり，1.0 よりも小さいとき，良好であると解釈する．このとき，シミュレーション回数は，500 回とする．因に，修正 RuleFit 法の Elastic Net 法のパラメータ (λ, η) は，事例検討と同様に，10 重交差確認法により推定する．

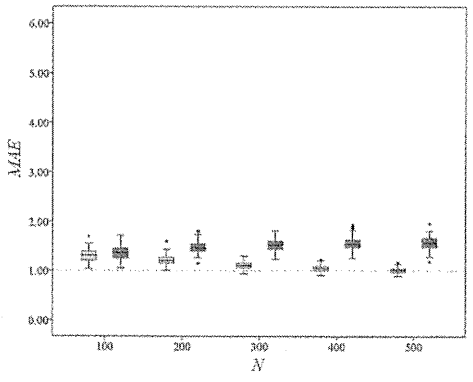
図 3 は，数値検証の結果である．真のモデルの線形傾向が強い場合において(図 3(a)(b)(c))，$RR = 6.00$(誤差分散が小さい)のとき，いずれのボックス・プロットのヒンジも 1.00 を含むことがなかった．すなわち，修正 RuleFit 法の性能が他の 2 手法に比して極端に良好な傾向を示した．標本サイズ N が増加するにつれて，MAE_{MART} が減少傾向を示したが，MAE_{RF} は増加傾向を示した(図 3(a))．したがって，RandomForest 法は，3 手法のなかで最も粗悪な結果だった．



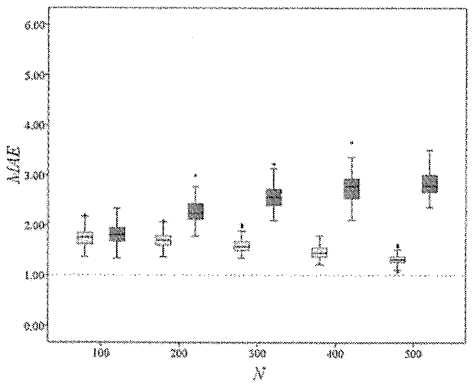
(a) $RL = 0.25,\ RR = 6.0$



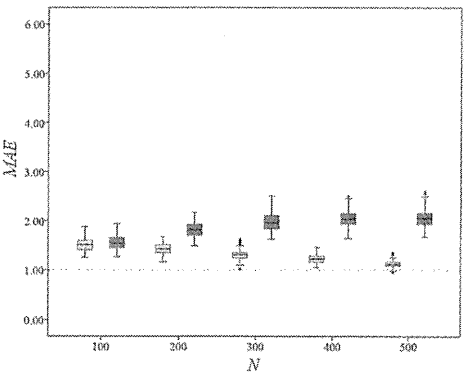
(b) $RL = 0.25,\ RR = 4.0$



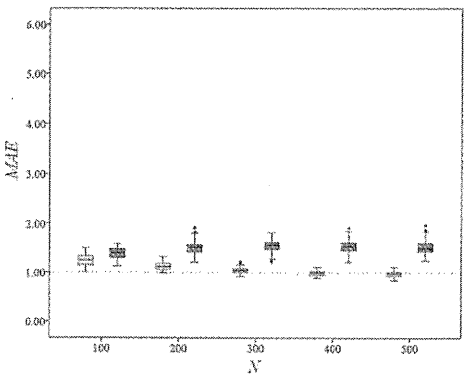
(c) $RL = 0.25,\ RR = 3.0$



(d) $RL = 1.00,\ RR = 6.0$



(e) $RL1.00,\ RR = 4.0$



(f) $RL = 1.00,\ RR = 3.0$

図 3. 数値検証の結果(濃い濃淡：RandomForest，薄い濃淡：MART)

Breiman (2001) および杉本 (2001) は，回帰問題において，RandomForest 法は，ブースティング接近法に比べて粗悪な結果を示すことを指摘しているが，本数値検証は，このことを裏付けた．RR が増加する(誤差分散が増加する)につれて，修正 RuleFit 法と他の 2 手法の差は小さくなっているものの，ボックスが 1.00 を下回ることはなかった(修正 RuleFit 法が他のアンサンブル学習法に劣ることがなかった)．修正 RuleFit 法は，基本学習器に線形項を含むため，RandomForest 法および MART 法よりも良好な結果を示すことが予想された．このとき，線形項を含むことの優位性は，誤差分散および標本サイズが小さくなるほど顕著だった．

真のモデルが非線形構造をもたない場合において(図 3(d)(e)(f))，$RR = 6.00$(誤差分散が小さい)とき，$RL = 0.25$ の場合と同様に，いずれのボックス・プロットのヒンジも 1.00 を含まなかった．ただし，図 3(a) に比べて，その差異は小さかった．線形構造をもつ場合と同様に，RandomForest 法と他の 2 手法の性能の差は，標本サイズが増加するにしたがって増加傾向を示した．RR が小さい場合(図 3(f))，修正 RuleFit 法は，標本サイズが $N \leq 200$ において，最も顕著な適切性が示唆されたが，RR が増加する(誤差分散が増加する)につれて，修正 RuleFit 法と MART の 2 手法の差は小さくなり，$N = 500$ のとき，MAE_{MART} の中央値が 1.0 を下回った．

図 3(h) において，$N = 500$ のときに，MAE_{MART} が僅かに 1.00 を下回った．したがって，誤差分散(ノイズ)が大きいとき，図 3(e)(f) では，$N = 100 \sim 200$ のときに，$MAE_{\mathrm{MART}} = 1.50$ 程度を示しており，その他の状況においても修正 RuleFit 法の性能が MART 法を上回った．すなわち，多くの場合に，修正 RuleFit 法の性能が他のアンサンブル法を上回った．

5. グラフィカル診断の方法

Friedman and Popescu (2008) は，RuleFit 法に対する諸種の統計量を提案している．一部の統計量に関しては，有効なグラフィカル表現の方法が示されているが，部分ルール重要度および部分変数重要度に関しては，その有用性に関する指摘は，殆ど行われておらず，グラフィカル表現の方法も提案されていない．ただし，若干の工夫により有効に利用できるように思われる．そのため，本節では，これらの統計量に対するグラフィカル表現の方法について触れる．

5.1. 部分ルール重要度プロット

Friedman and Popescu (2008) では，ルール(基本学習器)が応答に与える影響を診断する統計量として，ルール重要度および，観測値の部分集合に基づく部分ルール重要度を提案している．ただし，ルール重要度の有用性について指摘されているものの，部分ルール重要度の応用方法あるいはグラフィカル表現法については議論されていない．本節では，ルール重要度および部分重要度について触れたうえで，部分ルール重要度のグラフィカル表現の方法について述べる．

任意の標本 $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ に対するルール項 $r_k(\boldsymbol{x})$ のルール重要度は

$$RI_k(\boldsymbol{x}) = |\hat{\alpha}_k| \cdot |r_k(\boldsymbol{x}) - \varrho_k|, \tag{10}$$

で与えられる．また，線形項 $l_j(x_j)$ の場合には

$$RI_j(x_j) = |\hat{\beta}_j| \cdot |l_j(x_j) - \bar{l}_j|, \tag{11}$$

である．ここに，$\bar{l}_j$ は $l_j(x_j)$ の平均値である．

N 個の観測値 $\boldsymbol{x}_i, i = 1, \ldots, N$ の部分集合 D における部分ルール重要度は，式 (10) および (11) を用いることで，

$$RI_k(D) = \frac{1}{|D|} \sum_{x_i \in D} RI_k(\boldsymbol{x}_i), \quad RI_j(D) = \frac{1}{|D|} \sum_{x_i \in D} RI_j(x_{ij}), \tag{12}$$

で与えられる．ここに，$|D|$ は D に含まれる $\boldsymbol{x}_i$ の数である．

通常は，説明変数 $\boldsymbol{x}_i$ よりも，むしろ応答 y_i に関心をもつことが多い．そのため，応答 y_i の分位点の任意の範囲に対応する説明変数の部分ルール重要度が算出される．

RuleFit 法では，構成されるモデルが非線形でかつ交互作用を含むため，特定の部分集合(たとえば応答が大きい(あるいは小さい)観測値)に対して等しく影響を及ぼしているとは考えにくい．例えば，臨床研究において，ポジティブ・レスポンダー(有効性が顕著な患者像)に影響を与える要因(ルール)のすべてが，ネガティブ・レスポンダー(有効性が認められない患者像)と同じであることは少ない．このような場合に，式 (12) を応答の分位点の範囲で区切られた部分領域 D_q $(q = 1, 2, \ldots, Q)$ のもとで部分ルール重要度を計算し，これらをグラフィカルに表示することは，応答に対する基本学習器の影響の傾向変化を捉えることに繋がる．

本論文では，部分ルール重要度の応答の分位点に対する傾向変化を図示するために，変動ダイアグラム (Unbanek, 2002) を用いる．元来，変動ダイアグラムは，RandomForest 法において構成される個々の樹木での変数重要度を提示するために提案されている．他方，本論文では，横軸に応答のパーセント点，縦軸に任意のルールを表す格子を張り，その格子状に部分ルール重要度の大きさに対応する正方形を描写する．したがって，Unbanek (2002) での適用方法とは異なる．本論文では，これを部分ルール重要度プロットと呼ぶ．部分ルール重要度プロットのアルゴリズムを以下に示す：

PL.0: 初期設定：観測値の分割数 Q, 上位ルール数 C
PL.1: 観測値 $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ より，修正 RuleFit モデルを推定する $(\hat{F}_{\mathrm{EN-RFit}}(\boldsymbol{x}))$.
PL.2: 全観測値を用いてルール重要度 (12) を計算し，上位 C 個のルール $h_c(\boldsymbol{x})$, $c = 1, \ldots, C$ を抽出する．ここに $h_c(\boldsymbol{x})$ は，ルール項 (1)，あるいは線形項 (2) のいずれかの形式をとる．
PL.3: $\boldsymbol{x}$ を y の分位点に基づいて Q 分割し，部分集合 D_q, $q = 1, 2, \ldots, Q$ を得る．
PL.4: PL.2 で得られたルール $h_c(\boldsymbol{x})$ に対して，D_q での部分ルール重要度を式 (12) により計算する $(RI_c(D_q), c = 1, \ldots, C;\ q = 1, \ldots, Q)$.
PL.5: X 軸を分位点区分の番号 q，Y 軸を c 番目のルールとしたときの格子上で正方形の面積により $RI_c(D_q)$ の大きさを表現する．

部分ルール重要度プロットを用いることにより，ポジティブ・レスポンダーあるいはネガティブ・レスポンダーに対して影響を与えるルールの摘出および応答の分位点に対応したルールの影響の大きさの推移を省察できる．

5.2. 部分変数重要度プロット

Breiman et al. (1984) によって提案された，変数重要度は，応答に対する個々の説明変数の影響を精査するのに有用であることから，多くのアンサンブル学習法に実装されている．また，RuleFit 法に対しても同様に提案されている．ただし，Friedman and Popescu (2008) では，部

分ルール重要度と同様に，部分変数重要度(観測値の部分集合に基づく変数重要度)に関する有用なグラフィカル表現の方法が提案されていない．

本節では，RuleFit 法における変数重要度および部分変数重要度を略説した上で，部分ルール重要度のグラフィカル表現の方法について述べる．

修正 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，前節のルール重要度に基づいて定義される．そのため，修正 RuleFit 法では，任意の標本 $\boldsymbol{x}_i$ に対する変数重要度(部分変数重要度)が計算できる．

標本 $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ での変数 x_j における変数重要度 $VI_j(\boldsymbol{x})$ は

$$VI_j(\boldsymbol{x}) = RI_j(x_j) + \sum_{\boldsymbol{x}_j \in r_k} RI_k(\boldsymbol{x})/p_k, \tag{13}$$

で与えられる．ここに，$RI_j(x)$ は，線形項 x_j での部分ルール重要度 (11) であり，そして $RI_k(\boldsymbol{x})$ は，$\boldsymbol{x}$ を含むルール $r_k(\boldsymbol{x})$ の部分ルール重要度 (10) 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．したがって，式 (13) の第 2 項は，x_j を含むルール項でのルール重要度の平均値である．式 (13) をすべての観測値に対して計算し，その平均値をとれば，他手法でも頻用される変数重要度である．

また，N 個の観測値 $\boldsymbol{x}_i$, $i = 1, \ldots, N$ の部分集合 D での変数重要度 $VI_j(D)$ は

$$VI_j(D) = \frac{1}{N_D} \sum_{i \in D} VI_j(\boldsymbol{x}_i) \tag{14}$$

である．

本論文では，部分変数の重要度の傾向変化を図示する方法として，部分ルール重要度と同様に，変動ダイアグラム (Unbanek, 2002) を応用する．本論文では，この方法を部分変数重要度プロットと呼ぶ．

部分変数重要度プロットのアルゴリズムを以下に示す：

PV.0: 初期設定：観測値の分割数 Q

PV.1: 観測値 $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ より，修正 RuleFit モデルを推定する $(\hat{F}_{\mathrm{EN-RFit}}(\boldsymbol{x}))$.

PV.2: $\boldsymbol{x}$ を y の分位点に基づいて Q 分割し，部分集合 D_q を得る．

PV.3: PV.2 で得られた推定モデル $\hat{F}_{\mathrm{EN-RFit}}(\boldsymbol{x})$ を用いて，p 個の説明変数の部分従属度 (14) を，それぞれの部分集合に対して計算する $(VI_j(D_q),\ j = 1, \ldots, p,\ q = 1, \ldots, Q)$.

PV.4: X 軸を分位点区分の番号 q，Y 軸を変数 x_j の名称としたときの格子上で正方形の面積により $VI_j(D_q)$ の大きさを表現する．

部分変数重要度プロットを用いることで，ポジティブ・レスポンダーあるいはネガティブ・レスポンダーに対して影響を与える変数およびその強さ，あるいは応答の分位点に対応した変数の影響の大きさの推移を省察できる．

6. 医学データへの応用

本論文では，通常の RuleFit 法に対して，Elastic Net による若干の修正を行い，部分ルール重

表 3. 前立腺癌データに対するルール重要度

ルール	ルール重要度	
1	$(lcp \leq 0.36) \cap (lpsa \leq 3.24)$	100.0
2	linear :$lpsa$	54.1
3	$(lpsa \leq 3.25)$	41.3
4	$(lpsa \leq 2.24) \cap (lcp \leq 0.34)$	35.4
5	$(lweight > 3.13) \cap (lpsa \leq 2.44)$	24.3
6	$(lbph < -1.6) \cap (age > 50)$	23.4
7	$(lpsa > 0.95) \cap (lweight \leq 3.03) \cap (pgg45 > 8.00)$	22.3
8	$(lweight \leq 3.85) \cap (lbph > 0.50)$	21.1
9	$(lpsa > 0.41) \cap (lweight \leq 3.61) \cap (lcp > 1.65)$	20.7
10	$(lweight \leq 3.61) \cap (age > 58)$	19.4

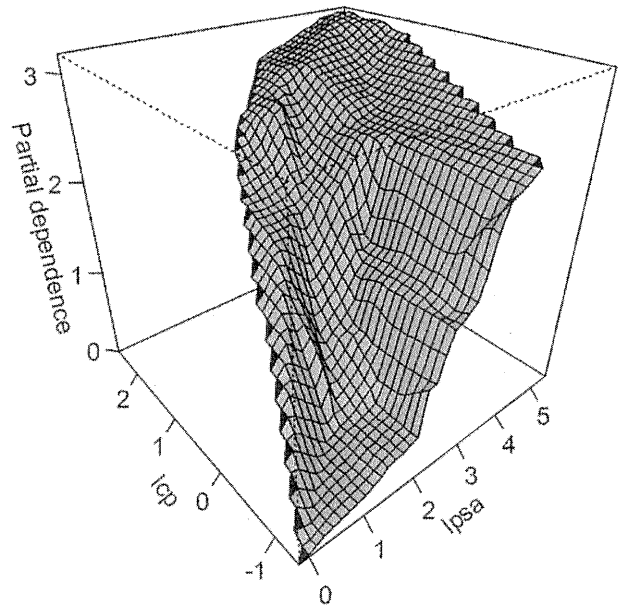


図 4. ***lcavol*** および ***lcp*** に対する 3 次元部分従属度 (Friedman, 2001)

要度および部分変数重要度に対するグラフィカル表現の方法を提示した．本節では，修正 RuleFit 法およびグラフィカル診断法の医学データへの応用について述べる．

いま，前立腺癌の腫瘍径と諸種の臨床的測度との関連性を評価するために，97 名の成人男性においてとられている (Stamey et al., 1989) を用いる．項目は，腫瘍径の対数値 ($lcavol$)，前立腺の重さの対数値 ($lweight$)，年齢 (age)，良性前立腺過形成量の対数値 ($lbph$)，精嚢の転移の有無 (svi)，莢膜の浸透率の対数値 (lcp)，Gleason スコア ($gleason$)，Gleason グレイドが 4 点あるいは 5 点の割合 ($pgg45$)，そして前立腺特異抗原の水準の対数値 ($lpsa$) である．本解析では，腫瘍径の対数値 ($lcavol$) を応答，他の 8 変数を説明変数とする．このとき，本解析の目標は，腫瘍径に対する影響要因を探索することにある．

ここでは，修正 RuleFit 法により構成される樹木数を 200，Elastic Net 法のパラメータの推定を η, λ を 10 重交差確認法によって選定した．このとき，$\hat{\eta} = 0.20, \hat{\lambda} = 0.018$ だった．

最も影響を及ぼす 10 個のルールとして，表 3 の結果が得られた．8 個のルールにおいて，前立腺特異抗原の水準の対数値 ($lpsa$) が含まれていた．また，莢膜の浸透率の対数値 (lcp) は，上位 5

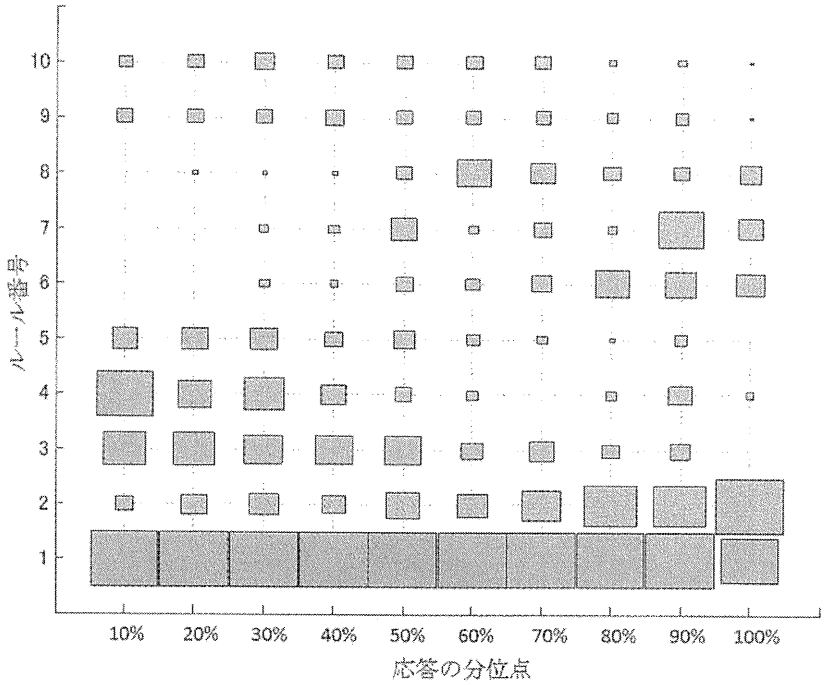


図 5. 前立腺癌データに対する部分ルール重要度プロット

個のルールのうち，3 個に含まれていた．さらに，ルール 1$(lcp \leq 0.36) \cap (lpsa \leq 3.24)$ およびルール 4$(lpsa \leq 2.24) \cap (lcp \leq 0.34)$ は，同じ変数により構成された．したがって，これらの変数の関連性の高さが伺える．図 4 は，$lpsa$ と lcp の 2 変量による部分従属プロット (Friedman, 2001) である．$lpsa$ および lcp の減少に伴い，腫瘍径の対数値 ($lcavol$) が減少傾向を示している．とくに，$lpsa$ と lcp の両方が小さい場合には，急激な減少傾向を示している．前立腺特異抗原とは，前立腺の細胞で生産される物質であり，前立腺癌のバイオマーカとして用いられている．また，莢膜の浸透率の上昇は，癌細胞の毒性の強さの増大を表す．前立腺特異抗原の水準の対数値および莢膜の浸透率といった癌細胞の活動の度合いに関連する要因が小さい場合に，腫瘍径に対するリスクは大幅に減少することが示唆される．

次いで，個々のリスク層の患者像 (腫瘍径の対数値 $lcavol$ の分位点毎での評価) に対する評価を行う．図 5 は，表 3 のルールに対する応答の 10%点刻みでの部分ルール重要度プロットである．その結果，ルール重要度が最も高かったルール $(lcp \leq 0.36) \cap (lpsa \leq 3.24)$ は，いずれの $lcavol$ においても，非常に高い部分ルール重要度を表した．2 番目に高かった $lpsa$ の線形項のルール重要度は，腫瘍径が大きくなるにつれて，部分ルール重要度が大きくなる傾向にあり，$lcavol$ の 90 パーセント点以上では，最大のルール重要度を示した．他方，3, 4 番目に高かったルール $(lpsa \leq 3.25)$, $(lpsa \leq 2.24) \cap (lcp \leq 0.34)$ では，腫瘍径が大きくなるにつれて減少傾向を示した．

図 6 は，$lpsa$ に対する部分従属プロット (Friedman, 2001) に対して，$lpsa$ の線形項を描写している．$lpsa$ が 3.5 付近よりも大きいとき，$lpsa$ の線形項付近に $lpsa$ に対する部分従属が布置しているものの，それ以下では，部分従属プロットが線形項 $lpsa$ に比べて極端に小さな値を示した．因みに，このときの $lpsa$ の回帰パラメータ 0.11 であった．急激な減少傾向を示した要因として，ルール 1$(lcp \leq 0.36) \cap (lpsa \leq 3.24)$ および，ルール 3$(lpsa \leq 3.25)$ の影響が考えられる．これらのルールに対する回帰パラメータの推定値は，それぞれ $-0.47, -0.32$ であり，かつ，サポート (全

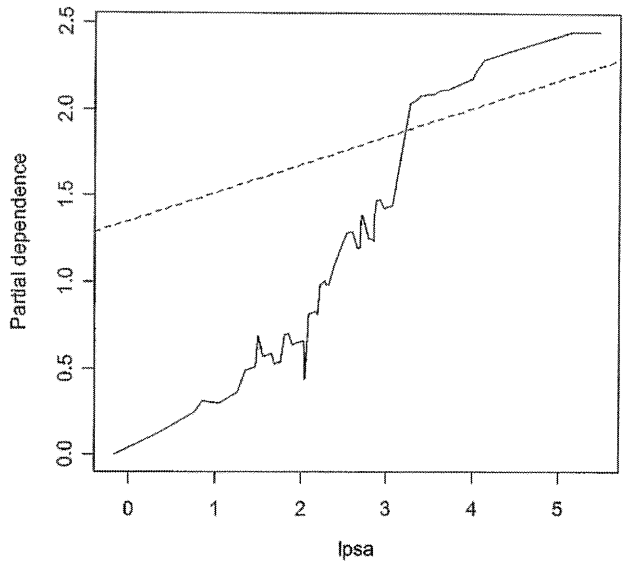


図 6. ***lcavol*** に対する部分従属度 (Friedman, 2001) および線形項 ***lcavol*** のプロット
(実線は部分従属度であり，点線は線形項 *lcavol* を表している)

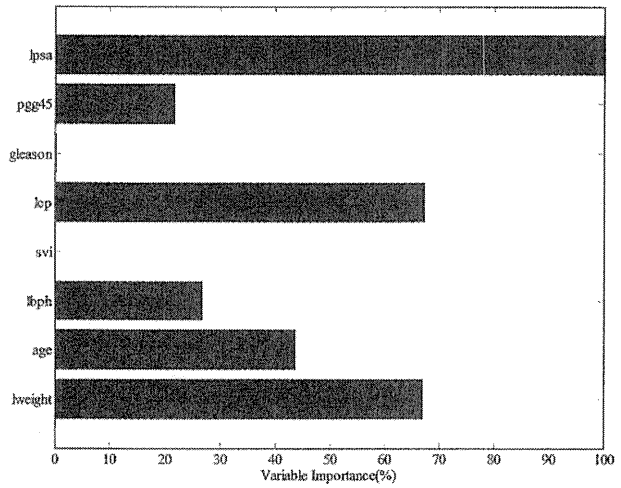


図 7. 前立腺癌データに対する変数重要度プロット

体に占めるルールに含まれる個体の割合)は，それぞれ，0.60, 0.76 であった(すなわち，多くの個体がこれらのルールによるリスク軽減が認められた)．したがって，高リスク群において *lpsa* の線形項の影響は顕著であるものの，ルール 3($lpsa \leq 3.25$) およびルール 4($lcp \leq 0.36$)$\cap$($lpsa \leq 3.24$) では，急激に腫瘍径に対するリスクを減少させた．($lcp \leq 0.36$)$\cap$($lpsa \leq 3.24$) による影響は，図 4 において考察したが，ルール 3($lpsa \leq 3.25$) に属する個体は，*lcp* の値に依らず，腫瘍径が 0.32 ほど小さくなる．

図 7 は，全観測値に対する変数重要度プロットである．ルール重要度の高いルールの多数に含まれていた *lpsa* の変数重要度が極端に高く，次いで，*lweight* および *lcp* が高かった．*lcp* は，先述したルール 1 あるいはルール 4 に含まれていることから，変数重要度が高くなることは平易に理解できる．他方，*lweight* は，表 3 のなかの半数(5 個)に含まれており，これにより，変数重要度が高くなったと推察される．また，ルール 6 およびルール 10 に含まれている *age* の変数重要度も比較的高かった．他方，*gleason* および *svi* の変数重要度は，ほぼ 0 であった．Gleason スコ

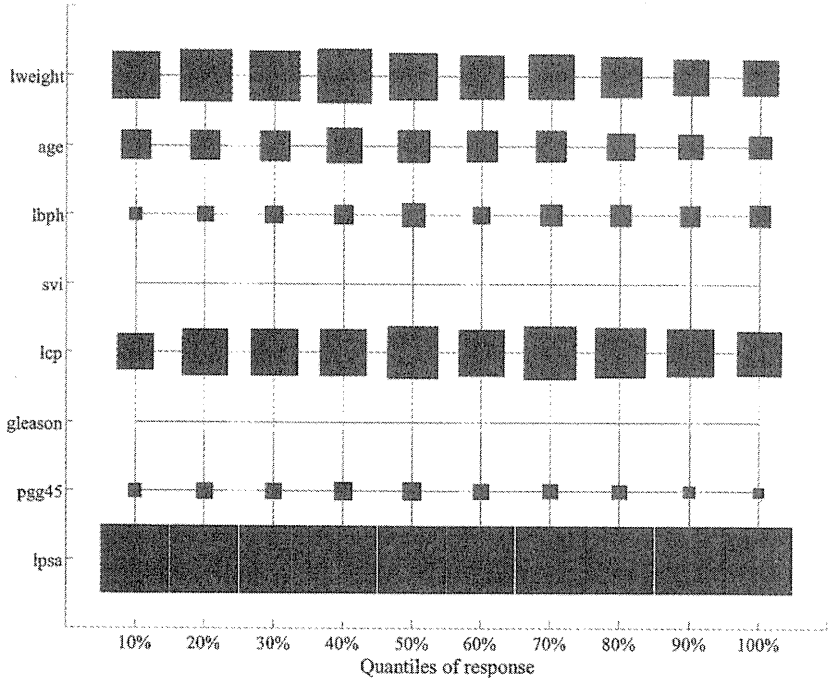


図 8. 前立腺癌データに対する部分変数重要度プロット

ア(*gleason*)は，前立腺癌の分化の程度を表す指標として，病理検査で広く用いられているが，その腫瘍径に対する影響が示唆されなかった．ただし，*pgg*45 の変数重要度が 26%程度存在することから，生検で採取した 2 個の大きな組織像のみに依存する Gleason スコアではなく，Gleason グレイドの高い組織像がどの程度存在するかが腫瘍径の大きさに対して重要なようである．また，精嚢(*svi*)は，前立腺の後ろ側にある器官であるが，精嚢への転移の有無によって腫瘍径が増減しないようである．

図 8 は，応答 *lcavol* の 10%点刻みでの部分変数重要度プロットである．*lpsa* の部分変数重要度は，いずれの応答の分位点においても最も高かった．*lweight* は，応答の分位点が高くなるにつれて，僅かな減少傾向を示した．しかしながら，顕著な傾向変化を与えることはなかった．

7. 結　　び

論文では，アンサンブル学習法の新たな接近法として，RuleFit 法を俎上にあげ，その修正版として Elastic Net 法による縮小回帰に基づく修正 RuleFit 法を提案した．さらに，その結果をグラフィカルに診断するためのグラフィクスを構成した．

本節では，本論文で得られた知見を要約し，結びに代える：

数値検証による修正 RuleFit 法の評価

A1. 真のモデルの線形傾向が強い場合には，通常 RuleFit 法は良好な性能を示したが，非線形傾向あるいはノイズが増加するにつれて，通常 RuleFit 法では，過剰刈り込みの傾向を示し，平均絶対偏差が増加傾向を示した．他方，修正 RuleFit 法は，このような状況において，より良好な結果を示した．

A2. 真のモデルの非線形構造の傾向が強く，かつ標本サイズが大きい場合には，修正 RuleFit

法に比して MART 法のほうが僅かに良好な性能を示した．ただし，標本サイズが減少するにつれて，修正 RuleFit 法が MART 法の性能を大幅に上回った．

A3. 真のモデルの線形傾向が強くなるほど，修正 RuleFit 法が MART 法および RandomForest 法に比して良好な性能を示した．また，ノイズが増加するにつれてその傾向は顕著になった．

修正 RuleFit 法およびグラフィカル診断による文献事例の解剖

B1. 腫瘍径に対して，最も顕著な影響を与えたのは，前立腺特異抗原であることが示唆された．前立腺特異抗原の対数値が 3.25 以下において，腫瘍径は大幅に減少し，さらに，莢膜の浸透率の対数値が 2.24 以下では，その減少傾向が増大した．他方，その 3.25 を超える場合には，線形的な増加傾向をもつことが示唆された．このとき，部分ルール重要度プロットは，個々のルール(および線形項)を解釈するのに有用だった．したがって，部分ルール重要度プロットにより，詳細に個々のルールを検討することは，部分従属プロットにのみに依存する既存のアンサンブル学習法での主観的な解釈に対して，ルール項に基づく客観的な形式で，より詳細な傾向変化の検討を行える．

B2. Gleason スコアあるいは精嚢が，腫瘍径の大きさに影響を及ぼさないことが変数重要度プロットによって示唆された．ただし，部分変数重要度プロットでは，顕著な傾向変化が省察されなかった．変数重要度プロットおよび部分変数重要度プロットは，推定されたモデル全体における個々の説明変数の応答に対する影響の大きさを示している．高い変数重要度を示す場合には，部分従属度による省察が推奨される．他方，部分変数重要度プロットでは，応答の部分集合に焦点を当て，その傾向変化を捉えている．このとき，極端な傾向変化，例えば，高リスク集合において極端に部分変数重要度が高くなる場合には，部分ルール重要度あるいは，推定された修正 RuleFit のモデルのルール(あるいは線形項)を精査することで，その要因を探索できると思われる．ただし，本データでは，このような極端な傾向変化を見出すことはできなかった．

謝　辞　丁寧な査読と不備な点のご指摘を頂戴した，本論文の審査員の先生方に深甚の謝意を捧げます．ありがとうございました．なお，本研究は，文部科学省私立大学戦略的研究基盤形成支援事業「セキュアライフ創出のための安全知循環ネットワークに関する研究(研究代表者：堀　雅洋(関西大学))」の支援を受けて行われた．ここに御礼申し上げます．

参　考　文　献

Breiman, L. (1996): Bagging predictors. *Machine Learning* **26**(2), 123–140.

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

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

Bühlman, P. and Yu, B. (2003): Boosting with L_2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98**, 324–338.

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

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

Friedman, J.H. and Fisher, N.I. (1999): Bump hunting in high dimensional data, *Statistics in Computing* **9**, 123–143.

Friedman, J.H., and Popescu, B.E. (2003): Importance Sampled Learning Ensembles. *Stanford University, Department of Statistics. Technical Report.*

Friedman, J.H., and Popescu, B.E. (2004): Gradient directed regularization for linear regression and classification. *Stanford University, Department of Statistics. Technical Report.*

Friedman, J.H., and Popescu, B.E. (2008): Predictive learning via rule ensemble. *Ann. Appl. Stat.* **2**(3), 916–954.

Harrison, D. and Rubinfeld, D.L. (1978): Hedonic prices and the demand for clean air. *Environ, Economics and Management* **5**, 81–102.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2009): *The Elements of Statistical Learning: Data mining, inference and prediction (2nd edition)*. Springer.

Morgan, J.N. and Sonquist, J.F. (1963): Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* **58**, 415–434.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989): Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients, *Journal of Urology* **16**, 1076–1083.

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

Urbanek, S. (2002): Exploring statistical forests, *Proceedings of the 2002 Joint Statistical Meeting*, http://simon.urbanek.info/simon/yawe/research/pub.html

Zou and Hastie, T. (2005): Regularization and variable selection via the elastic net. *J. Royal. Statist. Soc.* B. **67**(2), 301–320.

下川敏雄 (2010): ルール・アンサンブル法とその拡張, 大分統計談話会 第 41 回大会 発表抄録.

下川敏雄, 後藤昌司 (2010): 拡張型ルール・アンサンブル接近法とその診断, 日本計算機統計学会 第 24 回大会 論文集 45–48.

下川敏雄, 武藤由香里, 御園生拓, 北村眞一 (2009): アンサンブル学習法の河川景観満足度調査データ解析への応用, 景観・デザイン論文集 6, 41–50.

下川敏雄, 大山 勲, 風間ふたば, 西山志保, 北村眞一 (2010): 2 値応答に対する縮小推定型多重加法型回帰樹木の開発：水道水満足度への応用. 感性工学 9(4), 653–661.

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 18(2), 123–164.

(2010 年 8 月 5 日受付　2011 年 1 月 10 日最終修正　1 月 13 日採択)

著者連絡先：〒 400–8511 山梨県甲府市武田 4–3–11
山梨大学 大学院医学工学総合研究部　下川敏雄
Tel. 055–220–8395
E-mail: shimokawa@yamanashi.ac.jp

Japanese J. Appl. Statist. 40 (1) (2011), 19–40

Modified Rule Ensemble Method and its Application for Bioceutical Data

Toshio Shimokawa[1,*], Mitsuhiro Tsuji[2] and Masashi Goto[3]

[1] Graduate School of Medicine and Engineering, University of Yamanashi
[2] Faculty of Informatics, Kansai University
[3] Biostatistical Research Association, NPO

Abstract

Ensemble learning methods can improve the prediction accuracy by combining multiple base learners, and are studied in the fields of statistics science and data mining. Since ensemble learning methods construct models of a "black box" nature, the models are difficult to interpret. Friedman and Popescu (2008) proposed the rule ensemble learning method, in which nodes of tree models are used as base learners. The rule ensemble method not only presents the base learner as a production rule, but also gives the response variable an influential measure with rule importance. In the rule ensemble method, base learners are weighted by shrinkage regression using the least absolute shrinkage and selection operator (lasso). However, when some pairs of base learners have high correlation, the lasso method prunes base learners excessively. In this study, we utilized elastic net (Zou and Hastie, 2006) for weighting the base learner to solve the problem of excessive pruning. We called our rule ensemble method the EN-RF method. Furthermore, we developed diagnostic graphics for partial variable importance and partial rule importance. The usefulness of the EN-RF method and its diagnostic graphics were illustrated by a practical example in medical data. In application of medical data, we focused on the characterization of the positive (and/or negative) responder. We found that the EN-RF method shows better performance compared with the existing regression method.

Key words: ensemble learning, regression, elastic net, graphical diagnostic method

*Corresponding author
E-mail address: shimokawa@yamanashi.ac.jp (Toshio Shimokawa)

Received August 5, 2010; Received in final form January 10, 2011; Accepted January 13, 2011.

Behaviormetrika
Vol.38, No.2, 2011, 125–138

HIERARCHICAL CLUSTER ANALYSIS FOR MULTI-SAMPLE COMPARISONS BASED ON THE POWER-NORMAL DISTRIBUTION

Toshio Shimokawa* and Masashi Goto**

To make multi-sample comparisons in comparative experimental studies, multiple comparison methods are generally used. The primary aim of these methods is to test hypotheses for pairwise equality of means, but it is often difficult to extract particular features from the data. An alternative approach is to use cluster analysis to group the sample means, and then to categorize the sample means as significantly different if and only if they belong to different groups. This approach does not involve hypothesis tests for pairwise equality of means and provides a useful interpretation of sample difference based on a graphical display. In many clustering methods for multi-sample comparisons, the normality of the observations is assumed. However, real-world observations rarely satisfy this strict assumption. We therefore propose the power-normal multi-samples cluster analysis (PMC) method that assumes the distribution of the observations is power-normal. Here, the power-normal distribution is defined as the distribution before the power-normal transformation (Box and Cox, 1964). We illustrate the usefulness of the PMC method for ordinary cluster analysis for multi-sample comparisons by analyzing an example and by evaluating a small-scale simulation.

1. Introduction

In engineering or biosciences, it is often necessary to extract features that differ between two or more samples. To examine the difference between samples, we apply comparative methods, especially multiple comparison methods. The primary goal of these methods is to test hypotheses for pairwise equality of sample means. However, it is not enough only to know whether or not the treatment means are equal. There also lies interest in learning which of the treatment subsets have large means, and also in the way treatments are partitioned into groups (within which mean values are nearly equal but between which mean values are different).

However, statistical graphics can give useful indications regarding the difference between samples. For example, Tukey (1977) points out the usefulness of statistical graphics in multi-sample data analysis. He recommends clarifying the features of multi-sample data by grouping (clustering) samples. In the present study, we use a clustering method to perform multiple comparisons. In particular, our approach does not involve hypothesis tests for pairwise equality of means, nor do we need to adjust the level of significance, which depends on the number of groups. We review these statistical graphics, and call our approach the "cluster analysis for multi-sample comparisons."

Key Words and Phrases: multiple-comparison, experimental design, cluster analysis, Akaike's information criteria, power-normal distribution

* Graduate School of Medicine and Engineering, University of Yamanashi 4–3–11 Takeda, Kofu City 400–8511 Japan (Tel: +81-55-220-8395, E-mail: shimokawa@yamanashi.ac.jp)

** Biostatistical Research Association, NPO

In many cluster analyses for multi-sample comparisons, the observations are assumed to be normally distributed. However, real-world observations rarely satisfy this strict assumption. Worsley (1977) proposed Scott and Knott's nonparametric version (Scott and Knott, 1974). He used Kruskal-Wallis test statistics as dissimilarly measure. On the other hand, by using a nonparametric method it is possible to reduce the information that is inherent in data (for example, position, dispersion, and shape). Thus, we propose the power-normal multi-samples cluster analysis (PMC) method that assumes that the distribution of the observations is power-normal (Goto *et al.*, 1979, 1983; Goto and Inoue, 1980), where the power-normal distribution is defined as the distribution specified before the power-normal transformation (Boxand Cox, 1964).

In section 2, we review some existing methods for multi-sample clustering. In section 3, we propose the PMC and its diagnostic methods. In section 4, we evaluate their performance by some examples and simulations. In particular, we discuss the usefulness of PMC, with their diagnostic methods, for multi-sample comparisons, and we make a thorough comparison with available methods. Finally, we conclude with some remarks and discuss further developments.

2. Cluster analysis for multi-sample comparisons

In this section, we investigate the cluster analysis for multi-sample comparisons. In particular, we illustrate the hierarchical agglomerative cluster analysis for multi-sample comparisons based on Akaike's Information Criteria (AIC) by Bozdogan (1985) because the new method we propose in section 3extends this method.

2.1 Notation

Assume a set of size I_k from k populations with $\{x_{ki}\}_{i=1}^{I_k}$. If the sample means for each sample are $\{\bar{x}_k\}_{k=1}^{K}$ and the sample variances are $\{\hat{\sigma}_k^2\}_{k=1}^{K}$, then the average of k sample means is $\bar{x} = \sum_{k=1}^{K} \bar{x}_k/K$ and the pooled estimate of common variance σ^2 is $\sigma^2 = \sum_{k=1}^{K}\sum_{i=1}^{I_k} I_k^{-1}(x_{ki} - \bar{x})^2$. By definition, the clusters Q_n do not overlap each other, therefore, $k \in Q_n$, $Q_n \neq Q_1 \cup Q_2 \cup \cdots \cup Q_N = \{\text{allset}\}$, $n = 1, 2, \ldots, N$.

2.2 Ordinary methods

In the analysis of multi-sample data, we are interested in which sample subsets have the larger means and in the manner in which samples are clustered into groups (within which mean values are nearly equal but between which mean values are different). To detect significant differences between the clusters or groups, multiple comparison methods have generally been used. The primary aim of these methods is to test hypotheses for pairwise equality of sample means, but these methods often make it difficult to summarize and explain features of the results. An alternative approach is to use a clustering method to put any similarity measure into groups (Tasaki

et al., 1987).

The cluster analysis for multi-sample comparisons is based on the same framework as the ordinary clustering method; i.e., it can be classified into hierarchical and non-hierarchical cluster analyses for multi-sample comparisons. Cox and Spjøtvoll (1982) propose a non-hierarchical cluster analysis for multi-sample comparisons whose F-test p-value is dissimilar in one-way analysis of variance (ANOVA). Calinski and Corsten (1985) use the simultaneous F test by Gabriel (1964) for dissimilarity.

We classify hierarchical cluster analyses for multi-sample comparisons into divided and agglomerative types. For the divided type, we build a single cluster that contains all samples, and we divide the two clusters and/or samples with the maximum dissimilarity. We divide until the p-values of all clusters become significant. Scott and Knott (1974) use likelihood-ratio test statistics about an average equality.

In an agglomerative type, we build K clusters (i.e., all samples) and amalgamate the two clusters or samples with the minimum dissimilarity. Bozdogan (1986) proposes the agglomerative, hierarchical, cluster analysis for multi-sample comparisons using Akaike's Information Criteria (AIC) for dissimilarity, and Jolliffe (1975) applies the p-value of Newman–Keuls's multiple range test. Furthermore, Calinski and Corsten (1985) recommend using the p-value in the studentized range test for dissimilarity.

In the present article, we focus on the hierarchical agglomerative clustering method based on AIC. The reasons for this choice are multiple. First, compared to other methods, it is comparatively easier to extend this method. In addition, this method can be applied even if sample sizes differ in each sample. Finally, this method can be represented using a dendrogram in a clustering result.

2.3 Cluster analysis for multi-sample comparisons using Akaike's Information Criteria

Bozdogan (1986) has applied AIC to agglomerative measurements. We call this clustering technique as the B-method. The original B-method is constructed for multivariate normal distributed data. However, we focus on analyzing univariate. Therefore, we explain the B-method under univariate settings. In the present work, AIC is defined as

$$AIC = -2\{\text{maximum likelihood}\} + 2p,$$

where p is the number of parameters ($p = N + 1$) in the model. Let $\{x_{ik}\}_{i=1}^{I_K}(k = 1, 2, \ldots, K)$ be the kth sample with the normal distribution $\mathrm{N}(\mu_k, \sigma)$. For the N-partition, multi-sample cluster model, the AIC is expressed as

$$AIC_{\mathrm{ND}} = I\log(2\pi) + I\log(\hat{\sigma}^2) + I + 2(N+1), \tag{2.1}$$

where,

$$\tilde{x}_n = \left\{ \sum_{k\in Q_n} \sum_{i=1}^{I_k} x_{ki} \right\}, \quad N\hat{\sigma}^2 = \sum_{k=1}^{K}\sum_{i=1}^{I_k}(x_{ki} - \bar{x}_k)^2 + \sum_{n=1}^{N}\sum_{k\in Q_n}(\bar{x}_k - \tilde{x}_n)^2.$$

In the B-method, all possible agglomerations are enumerated and the pair with the

minimum AIC statistics is chosen as the optimal agglomeration. The algorithm is constructed as follows:

Bl: Take the initial value $N = k$ and estimate the common variance $\hat{\sigma}^2$ using all samples.

B2: Search for ${}_N C_2$ pairs obtained from N clusters and process the following two steps.

B2a: Calculate

$$\tilde{x}_{n_1 n_2} = \sum_{k \in (Q_{n_1}, Q_{n2})} \sum_{i=1}^{I_k} x_{ki} \Big/ \sum_{k \in (Q_{n_1}, Q_{n2})} I_k$$

for the cluster pair (n_1, n_2).

B2b: Calculate the AIC (equation (2.1)) for each pair.

B3: Amalgamate the cluster pair (n_1^*, n_2^*) with the minimum AIC as the optimal agglomeration $(N \to N-1)$.

B4: Continue steps B2 and B3 until $N = 1$.

3. Cluster analysis for multi-sample comparisons based on power-normal distribution

In this section, we present the agglomerative hierarchical cluster analysis for multi-sample comparisons based on the power-normal distribution. Furthermore, we describe the relationships between the B-method and the proposed method.

3.1 The power normal distribution

For a positive random value x, the power-transformation is defined as

$$x^{(\lambda)} = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & , \lambda \neq 0, \\ \log x & , \lambda = 0, \end{cases} \tag{3.1}$$

where λ is called the power-transforming parameter (Box and Cox, 1964), and $x^{(\lambda)}$ is a power-transformed observation of x. Next, the power-normal distribution is defined as the distribution that specifies the observation x before the power-transformation, at which point $x^{(\lambda)}$ is approximately normally distributed. The probability density function $f_{\mathrm{PND}}(x; \lambda, \mu, \sigma)$ is given by

$$f_{\mathrm{PND}}(x; \lambda, \mu, \sigma) = \frac{x^{\lambda-1}}{A(\lambda)\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x^{(\lambda)} - \mu)^2}{2\sigma^2} \right\}, \tag{3.2}$$

where μ and σ^2 are location and scale parameters, respectively, and $A(\lambda)$ is the probability proportional term of the power-normal distribution and is given by

$$A(\lambda) = \begin{cases} \Phi\left[\mathrm{sign}(\lambda)|\Xi|\right], & \lambda \neq 0, \\ 1, & \lambda = 0, \end{cases}$$