Vol.2011-BIO-25 No.8
2011/6/23

生存時間研究におけるルールアンサンブル法の開発

下川敏雄[†1]　辻　光宏[†2]

生存時間解析において，予後 (生存期間) の予測，および予後因子の探索は，重要な要件の一つである．とくに，抗がん剤研究では，レスポンダー探索が必須である．このような状況において，樹木構造接近法が広範に普及しつつある．ただし，樹木構造接近法の予測確度が低いことは，広く知られており，適切な結果を導かない惧れがある．

本報告では，ルール・アンサンブル法を生存時間解析に拡張することで，予測確度の高いモデルの構築を試みる．また，ルール・アンサンブル法では，lasso 法による縮小回帰モデルのパラメータを用いることで，基本学習器 (ルール) の重要度を評価できる．このことは，予後因子の順位付けに繋がるだけでなく，レスポンダー探索のための道標を提示することができる．

Development of rule emsumble method for survival data

Tosio Shimokawa[†1] and Mitsuhiro Tsuji[†2]

One of the important themes in survival analysis is to explore prognoses factors that influence survival time. Recently, the tree-structured method has been applied to evaluate covariates; however, it is well known that this method has provides poor prediction model. This problem could be improved by modeling many trees in a linear combination, namely, ensemble learning. The ensemble learning method is actively studied in machine learning and statistics. In this presentation, we extended the rule ensumble method to analyze survival data, namely survival rule fit method (SRF method). SRF model is constructed by Cox proportional hazard model, and weight (regression) parameters for each rule (base learner) are estimated by lasso.

†1 山梨大学 大学院医学工学総合研究部
Graduate School of Medicine and Engineering, University of Yamanashi
†2 関西大学 総合情報学部
Faculty of Informatics, Kansai University

1. はじめに

生存時間研究では，患者の予後要因の探索が重要な課題の一つである．ただし，その要因が個別に影響を与えることは少なく，多くの場合に，それらの交互作用が予後を左右する．このとき，予後要因の交互作用効果を捉えるための有用な道具が樹木構造接近法である．生存時間研究における樹木構造接近法 (以下，生存時間樹木構造接近法) は，(1) 予め，説明変数と応答の間の関数関係および応答の分布 (生存時間分布) を想定しなくてもよい (パラメトリック接近法で設定される仮定を必要としない) こと，(2) 説明変数の応答への寄与あるいは説明変数間の交互作用構造を把握することができることから，特定の疾病での生存時間に絡む影響要因 (共変量) を評価するための強力な道具として広まってきている[4]．

生存時間研究における樹木構造接近法には，2 種類の接近法が提案されている．一つは，検定統計量およびその p 値を分岐点に用いる接近法である．この方法では，多分岐による樹木の開発が行われている．もう一つは，分類回帰樹木法 (CART 法)[1] のアルゴリズムを発展させるために，生存時間データに対するふし内不均一性測度を提案する方法である[15]．

ただし，いずれの樹木構造接近法においても，構成されたモデルの予測確度が低いことは広く知られており，場合によっては，線形モデル (生存時間研究においては，Cox 比例ハザード・モデル[3]) の性能を下回ることもある．このことに対処するための方法として，LeBranc & Crowley[11] は，CART 法のステップ関数を滑らかな打ち切りベキ乗既定関数に拡張した，多変量適応型回帰スプライン法 (MARS 法) を生存時間研究に拡張している．生存時間 MARS 法は，生存時間 CART 法よりも予測確度に優れているものの，外れ値および多重共線性の影響を受けるだけでなく，個々の予後要因 (生存期間に影響を及ぼす因子あるいは交互作用により表されるプロダクション・ルール) に対する評価をハザード比のみで評価できないため，CART 法ほど広く適用されていない．

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も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，この流れに沿って多くの樹木に基づくアンサンブル学習法が提案および応用されている[16]．生存時間研究におけるアンサンブル学習法として，Hothorn *et al.*[8] は，対数変換された生存時間に対する (中途打ち切りによる) 重み付き観測値に対するアンサンブル学習法を提案している．他方，Ishiwaran *et a.*[9] は，RandomForest 法を生存時間研究に拡張した生存時間 RandomForest 法 (以下，RSF 法) を提案している．そこでは，生存時間樹木構造接近

法[7] をブートストラップ標本に基づいてアンサンブルさせる方法を提案するだけでなく，累積ハザード関数 (および生存時間関数) の out-of-bag 推定の方法を提案している．さらに，Ridgeway[10] は，確率勾配ぶースティング法に基づく方法を提案している．この方法では，Cox 比例ハザード・モデルの共変量の項を一般化加法モデル[7] の枠組みで捉え，基底関数として樹木を用い，部分尤度およびその偏分残差を最小化するようにモデルを構築している．

ただし，いずれのアンサンブル学習法においても，MART 法での問題と同様に，予後要因に対する解釈に関する問題を解決できない．Friedman & Popescu[5] は，個々の基本学習器を定量的に評価できる，回帰問題に対するアンサンブル型学習法として，ルール・アンサンブル法を提案している．MART(Multiple Additive Regression Trees) 法[7] あるいは RandomForest 法[2] といったアンサンブル学習法では，CART 樹木を基本学習器に用いるのに対して，RuleFit 法は，CART 樹木によって得られる根幹ふし (CART 樹木における末端のふし，リーフとも呼ばれる) 以外の子ふし (CART 樹木における分岐を伴うふし，分岐ふしあるいはノードと呼ばれる) および線形回帰項をアンサンブル・モデルの基本学習器に用いる点で，既存のアンサンブル学習法と異なる．因に RuleFit 法の名前は，CART 樹木のふし (基本学習器) が「If ∼ Then」の形式で解釈できるプロダクション・ルールにより提示できることに由来する．

本報告では，ルール・アンサンブル法を生存時間解析に拡張することで，予測確度の高いモデルの構築を試みる．また，ルール・アンサンブル法では，lasso 法による縮小回帰モデルのパラメータを用いることで，基本学習器 (ルール) の重要度を評価できる．このことは，予後要因の順位付けに繋がるだけでなく，レスポンダー探索のための道標を提示することができる．

2. 生存時間解析の概要

生存時間解析では，応答 (個体) の何らかの推定値 (点推定値) を得ることよりも，むしろ，個々の標本に対する生存時間分布 (すなわち，生存期間および生存率) を推定することが重要になる．そのため，ここでは，生存時間解析の概要について触れる．

2.1 生存時間分布

生存時間データは，個体の死亡 (あるいは故障) までの時間を観測して得られる．観測される生存時間 t^* は生存時間 (確率変数)T の実現値 t であるか，あるいは中途打ち切りが生じたときには，その中途打ち切り (censoring) 時点の値である．本報告でとり扱う右側中途打ち切りとは，個体の登録時点はわかるものの，追跡期間中に死亡 (event) が起こらなかったことを表す．T を真の生存時間，$T_{\rm cens}$ を中途打ち切り時点とすれば，観測された生存時間データは，その確率変数

$$T^* = \min(T, T_{\rm cens})$$

の実現値 t^* である．さらに，$\delta = 1$ を死亡 (event) が確認できた症例，$\delta = 0$ を確認できなかった症例とする．すなわち，δ は

$$\delta = \begin{cases} 1 & , T^* \le T_{\rm cens} \\ 0 & , T^* > T \end{cases}$$

で表される指標関数である．以降では，(t^*, δ) を便宜上 (t, δ) で表す．

生存時間 T は確率変数であり，ある分布に従っている．この生存時間分布は，三つの等価な関数 (死亡密度関数，生存時間関数，(累積) ハザード関数) で特徴付けることができる．

死亡密度関数 (確率密度関数)$f(t)$ は，個体が区間 $t \le T < t + \Delta t$ 内で死亡する確率を表す．すなわち，

$$f(t) = \lim_{\Delta t \to 0} \Pr\{t \le T < t + \Delta t\}/\Delta t$$

である．

生存時間関数とは，少なくとも時点 $t(> 0)$ で生存する確率

$$S(t) = \Pr\{t \le T\} = 1 - F(x)$$

である．ここに $F(x)$ は累積分布関数 $F(t) = \int_0^t f(u)du$ である．

ハザード関数 $h(t)$ とは，個体が時点 t まで生存しているとの仮定のもとで，区間 $t \le T \le t + \Delta t$ で死亡する確率

$$h(t) = \lim_{\Delta t \to 0} \Pr\{t \le T < t + \Delta t | T \ge t\}/\Delta t$$

を表す．また，ハザード関数の時点 0 から時点 t までの累積関数として，累積ハザード関数

$$H(t) = \int_0^t h(u)du$$

が定義される．これらの関数は，

$$h(t) = f(t)/S(t)$$

の関係があり，1 つが決まれば，残りを求めることができる．

2.2 比例ハザード・モデル

予後要因を探索するための一般的な方法が，Cox の比例ハザード・モデルである[3]．比例

ハザード・モデルは形式

$$h(t,\boldsymbol{x}) = \exp(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}) \quad (1)$$

で与えられる．ここに，$h_0(t)$ は基線ハザード関数を表している．上記からもわかるように，比例ハザードモデルは，基線ハザード $h_0(t)$ に対して，$\exp(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x})$ による重みづけを行うようにモデルが構築される．このとき，比例ハザード・モデルによる生存率の記述は，

$$S(t,\boldsymbol{x}) = S_0(t)^{\exp(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x})}$$

のように与えられる．もし，x_j が (0,1) で表されるダミー変数で与えられているとき，対応する回帰係数の指数値 $\exp(\beta_j)$ は，ハザード比と呼ばれ，ダミー変数の有無によるリスクの高さを表す (1 よりも大きな場合，変数 x_j によるリスクが高くなり，1 よりも小さな場合，リスクが低くなる)．

比例ハザード・モデルにおける回帰係数 $\boldsymbol{\beta}$ の推定には，様々な接近法が提案されている．ここでは，最も一般的な部分尤度法について触れる．式 (1) の基線ハザード関数では，説明変数 $\boldsymbol{x}$ が含まれておらず，一方で，比例項 $\exp(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x})$ には，生存期間 t が含まれていない (また，予め潜在基礎ハザード関数に対する生存時間分布 (理論分布) を想定しなければ，パラメータは存在せず，Kaplan-Meier 法などの適用が可能である)．部分尤度法では，そのことに注目し，個別に尤度 (対数尤度) を最大化する方法である．

いま，データ集合 $\{\boldsymbol{x}_i, t_i, \delta_i\}_{i=1}^N$ が与えられているとする．ここに，$\boldsymbol{x}_i$ は，患者 i に対する $p\times 1$ 共変量ベクトル，t_i は生存期間，そして，δ_i は中途打ち切り指標 (0：生存，1：死亡) である．このとき，部分尤度 (生存期間にタイがない場合) は，

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left\{\frac{\exp(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_i)}{\sum_{j\in R(t_i)}\exp(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_j)}\right\}^{\delta_i}$$

で与えられる．部分尤度に基づくパラメータ推定では，上式の対数値 (対数部分尤度)

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{N}\delta_i\sum_{k=1}^{p}\beta_j x_{ji} - \sum_{i=1}^{N}\delta_i\log\left\{\sum_{j\in R(t_i)}\sum_{j=1}^{P}x_{ji}\right\}$$

を最大にするように与えられる．

3. ルール・アンサンブル法

本節では，回帰問題に対するルール・アンサンブル法について説明する．

3.1 ルール・アンサンブル法の概要

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) を連結することで構成される．このとき，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[5] は，解釈可能性をもたせたアンサンブル学習法として，ルール・アンサンブル (RuleFit) 法を提案している．RuleFit 法では，樹木の個々のふし (根幹ふしを除くすべての子ふし) により得られるプロダクション・ルール (ルール項) だけでなく，および線形項を基本学習器に用いる．そして，個々の基本学習器に対して，lasso 法による縮小回帰推定による重み付けを行う．これにより，不必要な基本学習器が要素が削除される．このことは，いいかえれば基本学習器の刈り込み過程に繋がる．

RuleFit 法によるルール項の構成には，様々なサイズの樹木を評価するために，CART 法[1] の成長過程のみを用いる．いま，説明変数 x_j がとることができる，すべての可能な値の集合を S_j とする ($x_j \in S_j$)．そして，樹木のふし k において，変数 x_j がとることができる，S_j の部分集合を s_{jk} とする ($s_{jk} \subseteq S_j$)．このとき，ふし k におけるルール r_k は

$$r_k(\boldsymbol{x}) = \prod_{j=1}^{p} I(x_j \in s_{jk}) \quad (2)$$

で与えられる．ここに，$I(\cdot)$ は，括弧内が真なら 1，偽なら 0 をとる指標関数である．s_{jk} は，樹木の分岐点により得られる．説明変数 x_j が計量値のとき，s_k を規定するためのプロダクション・ルールは，区間 $s_{jk} = (x_{jk}^-, x_{jk}^+]$ である．ここに，x_{jk}^-, x_{jk}^+ は，それぞれ，ルール項 r_k における変数 x_j のルールの下限と上限である．また，x_j が計数値の場合には，s_k は，カテゴリの部分集合である．すなわち，$r_k(\boldsymbol{x})$ は，s_{jk} の論理積として解釈できる．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．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

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

である．ここに，t_m は終結ふしの数である．M 回のアンサンブル過程は，MART 法[7] と同様に，ステイジワイズ過程の流儀で行われる．ステイジワイズ過程とは，基本学習器の構成過程において，1 度構成された CART 樹木に対して，以降の反復で調整を行わない方法

であり，各反復で同時に調整するステップワイズ過程とは異なる[18]．

既存のアンサンブル学習法では，樹木 (基本学習器) の終結ふし数 (あるいは樹木の深さ) を予め設定する．RuleFit 法における終結ふし数の選定には指数乱数を用いる．すなわち，m 番目のアンサンブルにおける樹木の終結ふし数 t_m は

$$t_m = 2 + \mathrm{floor}(\gamma)$$

である．ここに，γ は指数分布 $\mathrm{E}(1/(\bar{t}-2))$ に従う乱数であり，$\mathrm{floor}(\gamma)$ は γ 以下の最大の整数である．さらに，$\bar{t}(\geq 2)$ は M 回のアンサンブルに対する期待終結ふし数を表す．これにより，期待終結ふし数を小さくしたもとで，様々な深さの樹木が構成できる．

MART 法あるいは RandomForest 法では，基本学習器に樹木のみを用いるため，真のモデルが線形構造をもつ場合に，多くの基本学習器を要する惧れがある．RuelFit 法では，この問題に対処するために，lasso 法[13] による重みづけ (縮小回帰あてはめ) の前に，線形項を基本学習器に追加する．ただし，線形項を含めることは，外れ値に頑健な樹木の特徴を阻害するかもしれない．そのため，Friedman & Popescu[5] は，修正型線形項

$$l_j(x_j) = \min(\delta_j^+, \max(\delta_j^-, x_j)) \tag{3}$$

を用いている．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[5] は $q \simeq 0.025$ を推奨している．

したがって，RuleFit 法のモデルは，ルール項 (2) および線形項 (3) を用いることで

$$F_{\mathrm{RFit}}(\boldsymbol{x}) = \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(\boldsymbol{x}) + \sum_{j=1}^{p} \beta_j l_j(x_j) \tag{4}$$

で与えれる．すなわち，ルール項 $r_k(\boldsymbol{x})$ をルールに含まれるとき 1，含まれないとき 0 をとるダミー変数と見做すと，RuleFit 法のモデルは，他のアンサンブル学習法と同様に，基本学習器の線形結合で表すことができる．

式 (4) の回帰 (重み) パラメータ $\{\alpha_k\}_{k=0}^{K}$,$\{\beta_j\}_{j=1}^{p}$ は，lasso 法による縮小回帰推定を用いることで

$$(\{\hat{\alpha}_k\}_0^K, \{\hat{\beta}_j\}_1^p) = \\ e \underset{\{\alpha_k\}_{k=0}^K, \{\beta_j\}_{j=1}^p}{\mathrm{argmin}} \left[\sum_{i=1}^{N} L \left\{ y_i, \quad \alpha_0 + \sum_{k=1}^{K} \alpha_k r_k(\mathbf{x}_i) + \sum_{j=1}^{p} \beta_j l_j(x_{ij}) \quad + \lambda \cdot \left(\sum_{k=1}^{K} |\alpha_k| + \sum_{j=1}^{p} |\beta_j| \right) \right\} \right]$$

により得られる．ここに，$\lambda \geq 0$ は，lasso 型罰則に対する調整パラメータである．λ の最適パラメータは v 重交差確認法により推定できる．$L(\cdot)$ は，最小 2 乗損失などの何らかの損失関数である．

RuleFit 法による回帰パラメータ推定過程では，ルール項に対して，

$$sd_k = \sqrt{\varrho_k(1-\varrho_k)}, \tag{5}$$

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

$$\varrho_k = \frac{1}{N} \sum_{i=1}^{N} r_k(\boldsymbol{x}_i), \tag{6}$$

である．また，線形項に対する規準化は

$$l_j(x_j) \leftarrow 0.4 \cdot l_j(x_j)/\mathrm{std}(l_j(x_j)), \tag{7}$$

で与えられる．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 (6) が一様分布 U(0,1) に従うと仮定したときの標準偏差 (5) である．

3.2 ルール・アンサンブル法における諸種の統計量

ルール・アンサンブル法では，樹木に基づくアンサンブル学習法と異なり，基本学習器を個別にプロダクション・ルールにより解釈できる．ただし，基本学習器の数が膨大であることから，その解釈を支援するための諸種の統計量が提案されている．ここでは，ルール重要度，変数重要度，および部分従属度について触れる．

3.2.1 ルール重要度

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

$$RI_k(\boldsymbol{x}) = |\hat{\alpha}_k| \cdot |r_k(\boldsymbol{x}) - \varrho_k|, \tag{8}$$

で与えられる．また，線形項 $l_j(x_j)$ の場合には

$$RI_j(x_j) = |\hat{\beta}_j| \cdot |l_j(x_j) - \bar{l}_j|, \tag{9}$$

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

全データ集合 $\boldsymbol{x}_i, i = 1, \ldots, N$ では，データ集合全体に対して，式 (8)(9) を計算し，その平均値により与えられる．

3.2.2 変数重要度

Breiman *et al.*[1] 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．ルール・アンサンブル法では，lasso 法によって推定された個々の基本学習器に対する回帰係数

(重み) に基づいて定義される．

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

$$VI_j(\boldsymbol{x}) = RI_j(x_j) + \sum_{\boldsymbol{x}_j \in r_k} RI_k(\boldsymbol{x})/p_k, \tag{10}$$

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 (9) 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 (8) 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 (10) の第 2 項は，x_j を含むルール項でのルール重要度に平均値である．式 (10) 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．

全データ集合 $\boldsymbol{x}_i, i = 1, \ldots, N$ では，ルール重要度と同様に，データ集合全体に対して，式 (10) を計算し，その平均値により与えられる．

3.2.3 部分従属度

ルール・アンサンブル法では，他のアンサンブル学習法と同様に，推定されたモデルをそのままの形式では評価できない．このとき，Friedman[7] によって提案された，部分従属度の利用が有用である．

いま，データ集合 $\{(t_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^N$ (ここに，$\boldsymbol{x}_i = (x_{1i}, \cdots, x_{Pi})^{\mathrm{T}}$ である) が与えられているとする．関心がある $P^+ (< P)$ 個 (視覚的に表現するためには，$P^+ = 1, 2$) の共変量を $\boldsymbol{x}_s$ とし，それ以外を $\boldsymbol{x}_{\setminus s}$ とする ($\boldsymbol{x}_s \cup \boldsymbol{x}_{\setminus s} = \boldsymbol{x}$)．

このとき，部分従属度の経験推定値は，

$$\widehat{PD}_s(\boldsymbol{x}_s) = \frac{1}{N} \sum_{i=1}^{N} \hat{f}_{\mathrm{RFit}}(t|\boldsymbol{x}_s, \boldsymbol{x}_{\setminus s}) \tag{11}$$

で与えられる．

4. 生存時間ルール・アンサンブル法

生存時間ルール・アンサンブル法のモデルは，式 (1) の類推に基づき，

$$h(t; \boldsymbol{x}) = h(t) \exp\left(\sum_{k=1}^{K} \alpha_k r_k(\boldsymbol{x}) \right) \tag{12}$$

で与えられる．ここに，$r_k(\boldsymbol{x})$ は，ルール項である．回帰モデルに対するルール・アンサンブル法のモデル (4) では，線形項が含まれているものの，本報告では，含めないこととする．これは，生存時間研究では，予後要因によるハザード比を評価することが多いためである．

このとき，ルール・アンサンブル法を生存時間データに拡張するには，(1) 樹木による生存時間アンサンブル・モデルの構成，(2) 基本学習器に対するパラメータ推定，が重要になる．他方，これらが解決できれば，4 節で述べた回帰問題に対するルール・アンサンブル法のアルゴリズムをそのままの形式で利用できる．

樹木による生存時間アンサンブル・モデルの構成には，Ridgeway[10] の一般化 Boosting 法を用いる．この方法は，Friedman[7] によって提案された，確率勾配ブースティング法の損失関数を尤度に基づく偏分残差におき変えることで，諸種の応答に対するモデルを構築することを可能にしている．これは，樹木を基本学習器にした場合，得られる最終モデルは，それぞれの終結ふし (リーフ) に対応する (0,1) のダミー変数と捉えれば，通常の一般化線形モデルの枠組みで捉えることができることにある．

比例ハザード・モデルの部分尤度における偏分残差は，

$$Dev = -2 \sum_{i=1}^{N} \delta_i (f_{\mathrm{Bst}}(\boldsymbol{x}_i)) - \log(R_i), \quad \text{where} \quad f_{\mathrm{Bst}}(\boldsymbol{x}_i) = \sum_{m=1}^{M} \tau_m(\boldsymbol{x}_i) \tag{13}$$

で与えられる．ここに，R_i は，個体 i に対するリスク集合の個数であり，$\tau_m(\boldsymbol{x}_i)$ は，m 番目のアンサンブルにおける樹木を表している．

また，確率勾配 Boosting 法では，損失関数の 1 次導関数による疑似残差 (puseudo residual) に対して，それぞれの樹木が当てはめられる．個体 i に対する上式に対する疑似残差 $\tilde{e}_i$ は比例ハザード・モデルの部分尤度における偏分残差は，

$$\tilde{e}_i = \delta_i - \sum_j \delta_j \frac{I(t_i \geq t_j) \exp\{f_{\mathrm{Bst}}(\boldsymbol{x}_i)\}}{\sum_k I(t_k \geq t_j) \exp\{f_{\mathrm{Bst}}(\boldsymbol{x}_k)\}} \tag{14}$$

である．

得られたブースティング樹木 $\tau_m(\boldsymbol{x}), m = 1, \ldots, M$ からのルールの抽出は，通常のルール・フィット法と同様に，M 個の樹木の根幹ふし以外のふしに分け，ルール項 (および対応するダミー変数)$\tilde{x}_{k,i} = r_k(\boldsymbol{x})$ を構築する．

パラメータ α_k の推定には，比例ハザード・モデルに対する lasso 法[14] を用いればよい．このとき，生存時間 lasso 法による推定では，偏分残差に基づく条件付き最適化問題

Vol.2011-BIO-25 No.8
2011/6/23

$$\arg\min\{Dev(\boldsymbol{\alpha})\}, \quad \text{subject to} \quad \sum_{k=1}^{K} |\alpha_k| \leq s$$

を解くことで与えられる．

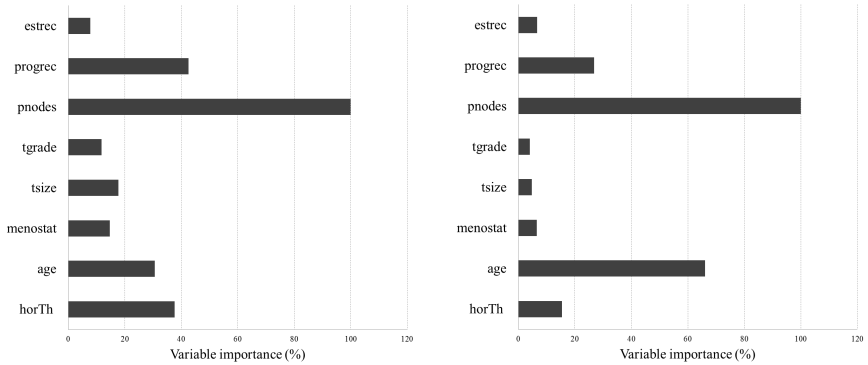
5. 事例検討

乳癌に罹患している患者に対して，ホルモン療法の効果を検討するために，ドイツ乳癌研究グループが 7 個の予後因子 (年齢 (age)，閉経の有無 (menostat)，腫瘍径 (size)，腫瘍のグレイド (grade)，リンパ節転移個数 (pnodes)，プロゲステロン・レセプタ個数 (progrec)，エストロゲン・レセプタ個数 (estrec)，およびホルモン療法が行われたか否か (hormone) とともに生存時間を評価している[12]．ここでの目標は，患者の生存時間に及ぼす共変量の影響 (とくに，ホルモン療法の影響) を評価することにある．

表 1 は，当てはめた生存時間ルール・アンサンブル法に基づいて，0.50 以上のルール重要度をもつルールの一覧を表している．9 個のルールのなかの 7 個のルールのなかに，リンパ節転移個数 (pnodes) が含まれた．リンパ節転移個数は，癌の進行程度を精査するうえで重要な要因の一つであり，乳癌においても，病期 (ステイジ) を分類するうえで用いられている．次いで，プロゲステロン・レセプタ個数 (progrec) を含むルールが 5 個存在し，また，ルール重要度が最も高いルールは，プロゲステロン・レセプタ個数 (progrec) によって構成された．プロゲステロン・レセプタは，乳癌における病理検査において調べられるホルモン・レセプターである．エストロゲン・レセプタ個数 (estrec) を伴うルールは得られなかった．また，ホルモン療法が行われたか否か (horTh) による影響では，腫瘍径が 7.5 超であり，プロゲステロン・レセプタ個数 (progrec) が 24 以下の場合に，ホルモン療法を実行しなければ，それ以外の場合に比べて，予後が悪くなることが示唆された．

表 1 ルール重要度が 0.5 以上のルールの一覧

ルール	ルール重要度	ハザード比	サポート
(progrec $>$ 21)	100.0	0.94	0.19
(pnodes $>$3) $\cap$ (progrec $\leq$ 55)	95.8	1.08	0.69
(pnodes $>$5) $\cap$ (progrec $\leq$ 8)	74.4	1.06	0.63
(pnodes $>$7) $\cap$ (progrec $\leq$ 24) $\cap$ (horTh=False)	66.7	1.11	0.05
(pnodes $\leq$ 5) $\cap$ (estrec $>$ 9.5) $\cap$ (age $>$ 54)	62.7	0.94	0.15
(pnode $>$ 3) $\cap$ (progrec $\leq$ 103.5)	57.5	1.01	0.08
(pnode $>$ 4) $\cap$ (tsize $>$ 27)	56.8	1.04	0.28
(pnodes $>$ 3)	56.7	1.04	0.55
(progrec $>$21)	53.8	0.98	0.25



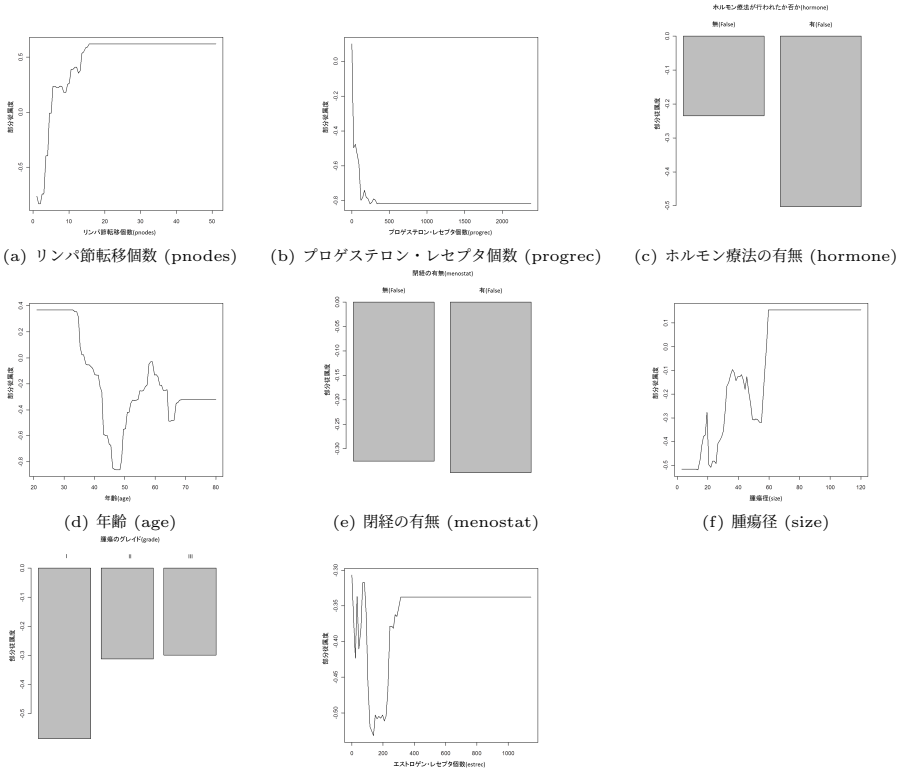
(a) 生存時間ルール・アンサンブル法　　(b) 生存時間ランダム・フォーレスト法

図 1 2 種類の生存時間アンサンブル学習法における変数重要度

図 1 は，変数重要度のプロットを表している．これは，個々の共変量が生存期間 (すなわち，ハザード) に及ぼす影響の大きさを評価するために用いられる．ここでは，生存時間ルール・アンサンブル法の結果とともに，対照手法として，生存時間ランダム・フォーレスト法の変数重要度を描写している．いずれの手法でも，リンパ節転移個数 (pnodes) の変数重要度が極端に高かった．生存時間ルール・アンサンブル法では，影響の強かったルールの多くに含まれていた．プロゲステロン・レセプタ個数 (progrec) での変数重要度が 2 番目に高かった．他方，生存時間ランダム・フォーレスト法では，年齢 (age) による影響が顕著に認められた．表 1 では，年齢を含むルールが 1 個だけであり，その結果，生存時間ルール・アンサンブル法では，年齢の影響は，それほど高くなかった．さらに，生存時間ルール・アンサンブル法では，ホルモン療法が行われたか否か (horTh) の影響が認められたが，生存時間ランダム・フォーレスト法では，それほど高くなかった．このとき，生存時間ルール・アンサンブル法での偏分残差は，0.287 だったのに対して，生存時間ランダム・フォレスト法では 0.394 だった．したがって，生存時間ルール・アンサンブル法は，生存時間ランダム・フォーレスト法に比べて良好な結果を示した．

図 2 は，生存時間ルール・アンサンブル法における部分従属度である．リンパ節転移個数 (図 2(a)) および腫瘍径 (図 2(f)) では，数値が増加するほど，部分従属度が増加した．したがって，リンパ節転移個数および腫瘍径の増加は，死亡リスクを増加させる傾向にあった．

Vol.2011-BIO-25 No.8
2011/6/23



(a) リンパ節転移個数 (pnodes)　(b) プロゲステロン・レセプタ個数 (progrec)　(c) ホルモン療法の有無 (hormone)

(d) 年齢 (age)　(e) 閉経の有無 (menostat)　(f) 腫瘍径 (size)

(g) 腫瘍のグレイド (grade)　(h) エストロゲン・レセプタ個数 (estrec)

図 2　個々の説明変数に対する部分従属度

リンパ節転移個数が増加することは，多臓器などへの転移などが推察されることから，その数値が増加することは，死亡リスクの増加に繋がることは，平易に理解できる．また，腫瘍径は，原発である乳癌の進行を表しており，これも同様に理解できる．プロゲステロン・レセプタ個数 (図 2(b)) では，数値が増加するほど，部分従属度 (死亡リスク) が急激な減少傾向を示したが，早い段階で，その傾向は飽和した．ホルモン療法の有無 (図 2(c)) では，ホルモン療法を施行したほうが，施行しない場合に比べて死亡リスクが軽減していることが顕著に認められた．他方，閉経の有無 (図 (e)) では，その差異が明確には認められなかった．年齢 (図 2(d)) では，55 歳付近まで急激な死亡リスクの減少が認められた後で，60 歳付近まで死亡リスクが増加傾向を示した．一般に，若年者での癌の進行は早いため，乳癌による死亡リスクが増加する傾向が認められ，また，高齢者では体力などの衰えによる死亡リスクの増加が懸念される．このことが，部分従属プロットに反映されたと思われる．腫瘍のグレイド (図 2(g)) では，最も数値の低いグレイド I での死亡リスクが極端に低いものの，グレイド II とグレイド III では，その差異が認められなかった．最後に，エストロゲン・レセプタ個数 (図 (h)) では，顕著な傾向変化が認められなかった．エストロゲン・レセプタ個数は，変数重要度でも，推定モデルに殆ど影響を与えないことが示唆されていることから，このことが反映されたと示唆される．

6. 結　　び

本報告では，アンサンブル学習法の新たな接近法として，ルール・アンサンブル法を俎上にあげ，生存時間データへの拡張を試みた．既存の生存時間アンサンブル学習法では，モデルが「ブラック・ボックス化」されるため，その結果の解釈は，変数重要度および部分従属度からの類推のみに終始した．他方，ルール・アンサンブル法では，基本学習器の重要度を示すことができるため，例えば，予後に影響を与える要因をプロダクション・ルールで捉えることができ，かつ，その影響 (リスクの高さ) の大きさを，lasso 法により推定された回帰係数から導かれるハザード比により評価できた．また，既存のアンサンブル学習法において提案されている諸種のグラフィカル接近法をそのままの形式で利用することができた．

謝辞　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

1) Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J.: *Classification and Regression Trees*. Chapman & Hall, (1984).
2) Breiman, L.: Random forests. *Machine Learning*, Vol.45, pp.5–32, (2001).
3) Cox, D.R. ：Regression models and life tables (with discussion), *J. Roy. Statist. Assoc.* Vol.B34, No.187–220, (1972).
4) Crowley, J., and Ankerst, D.P. : *Handbook of Statistics in Clinical Oncology*, Chapman & Hall, , (2006).
5) Friedman, J. H., and Popescu, B. E. : Predivtive Learning via rule ensemble. *Ann. Appl. Stat.*, Vol.2, No.3, pp.916–954 (2008).
6) Friedman, J. H. ：Multivariate adaptive regression splines, *Annals of Statistics*,

Vol.19, No.1, pp.1?-67.
7) Friedman, J. H.： Greedy function approximation: a gradient boosting machine.*Annals of Statistics*, Vol.29, pp.1189 –1232, (2001).
8) Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and van der Laan, M.J., Survival ensembles, *Biostat.*, Vol.7, pp.355–373, (2006).
9) Ishwaran, H., Udaya, U.B., Blackstone, E.H., and Lauer, M.S., Random survival forest, *Ann. Appl. Statist.*, Vol.2, No.3, pp.841–860, (2008).
10) Ridgeway, G.: Generalized boosted models : A guide to the gbm, available from ⟨http: //i-pensieri.com/gregr/papers/gbm-vignette.pdf⟩ (accessed 2008-6-6).
11) LeBlanc, M. and Crowley, J.：Adaptive regression splines in the Cox model, *Biometrics*, Vol.55, pp.204-213.
12) Schumacher, M., Basert, G., Bojar,H., Huebner,K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., and Neumann, R.L.A. : Rauschecker for the German Breast Cancer Study Group, Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive lbreast cancer patients,*Journal of Clinical Oncology*, Vol.12, pp.2086–2093, (1994).
13) Tibshirani, R. : Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.,* Vol,58, pp.267–288, (1996).
14) Tibshirani, R.: The lasso method for variable selection in the cox model, *Statistics in Medicine*, Vol.16, pp.185–395,(1997).
15) 衛藤俊寿・下川敏雄・後藤昌司：生存時間研究における多分岐型樹木構造接近法. 行動計量学, Vol.34, No.1, pp.1-20, (2007).
16) 下川敏雄・大山勲・風間ふたば・西山志保・北村眞一：2 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, Vol.9, No.4, pp.653-661, (2010).
17) 下川敏雄・辻光宏・後藤昌司：Elastic Net 罰則によるルールアンサンブル法とその応用, 応用統計学, Vol.40, No,1, pp.19-40, (2011).
18) 杉本知之・下川敏雄・後藤昌司：樹木構造接近法と最近の発展. 計算機統計学, Vol.18, No.2, pp.123–164, (2005).