競合リスクを伴う生存時間データに対するルール・アンサンブル法の開発

和歌山県立医科大学 臨床研究センター　下川敏雄

動機と目標

生存時間研究では，様々な原因によりイベントが発生する．このとき，関心のあるイベントがあるものの，そのイベント以前に異なるイベントが発生することで，調査不能になることが少なくない．関心のあるイベントと異なるイベントを競合リスク (competing risk) という．例えば，心血管系の患者に対する治療法を評価するための臨床試験において，全生存時間が主要評価項目とする．このとき，対象患者が癌により死亡した場合，心血管系イベントによる死亡までの生存期間は不明になる．このとき，癌による死亡が競合リスクになる．

本研究では，部分分布ハザード・モデルに焦点を当て，その予後要因を抽出するためのアンサンブル学習法を開発する．そこでは，基本学習器にプロダクション・ルールを用いるルール・アンサンブル法 (Friedman & Popescu, 2008) を適用する．これにより，関心のあるイベントに対する予後要因，競合リスクに対する予後要因をプロダクション・ルールによって提示することができる．

競合リスクを伴う生存時間ルール・アンサンブル法

競合リスクを伴う生存時間解析では，個々のイベントに対する生存時間分布よりも，むしろ累積発生関数 (CIF) を用いることが多い．また，競合リスクを伴う生存時間解析では，2 種類のハザード関数，すなわち，原因別ハザード関数と部分分布ハザード関数が存在する．原因別ハザード関数は，原因 j のそれぞれに対してハザード関数を計算する方法であり，競合リスクは中途打ち切りとみなされる．これに対して，部分分布ハザード関数では，競合リスクに基づいてリスク集合が修正される．

時間 t におけるイベント j の部分分布ハザード関数 $\gamma_j(t)$ は

$$\gamma_j(t) = \lim_{\Delta\to 0} \frac{\Pr\{t \le T < t+\Delta t, J = j | (T > t) \cup (T \le t \cap J \ne j\}}{\Delta t} = \frac{-d(\log(1 - I_j(t))}{dt}$$

で与えられる．すなわち，部分分布ハザード関数とは，原因 j によるイベントが時間 t までに発生しないか，あるいは競合リスクイベントが観測された条件のもとで次の時間 $t + \Delta t$ で発生する確率である (Grey, 1988)．

近年，競合リスクを伴う生存時間解析に対する幾つかの樹木構造接近法が提案されている．Callaghan(2008) は，ふし間分離測度 (LeBranc & Crowley, 1993) に基づく方法を提案している．そこでは，分離測度として Gray 検定の検定統計量を用い，LeBranc & Crowley(1993) のモデル構築過程に基づいて樹木を構成している．さらに，ふし内不均一性測度 (Breiman *et al.*,1984;LeBlanc & Crowley, 1993) に基づく方法では，原因別比例ハザード・モデルの Martingale 残差を利用している．また，アンサンブル学習法，とくにランダム・フォーレスト法の拡張では，Ishiwaran *et al.*(2014) が Gray 検定 (あるいは一般化ログランク検定) を用いる方法を提案しており，Mogensen & Gerds(2013) は，擬似値 (pseudo-value) を用いることで，通常のランダム・フォーレスト法をそのまま利用できることを指摘している．ただし，CART 法を拡張した方法では，予測確度が乏しいことは広く知られており，また，主効果の評価には不向きである．また，ランダム・フォーレスト法に代表されるアンサンブル学習法では，予測確度に優れているものの，その解釈は困難である．そのため，本報告では，予測確度に優れ，かつ推定されたモデルのなかで予測に対して影響の強い要因が「IF ∼ Then」のプロダクション・ルールで解釈できる RuleFit 法 (Friedman & Popescu, 2008) を俎上にあげた．

時間 t におけるイベント j の基線部分分布ハザード関数を $\gamma_0^j(t)$ とするとき，部分分布生存時間 RuleFit 法のモデルは，

$$\gamma^j_{\mathrm{SDRF}}(t|\boldsymbol{x}) = \gamma_0^j(t) \exp\left\{\sum_{k=1}^{K} \alpha_k^j r_k(\boldsymbol{x}) + \sum_{p=1}^{P} \beta_p^j l_p(x_p)\right\}$$

である．ここに，α_k^j はルール項 $r_k(\boldsymbol{x})$ に対するイベント j での回帰パラメータ，β_p は共変量 x_p のイベント j での修正線形項に対する回帰パラメータである．

RuleFit 法の構成は，(a) 樹木モデル (樹木に基づくアンサンブル学習法) に基づくルール項の推定，(b) 縮小推定法 (例えば，lasso 法) によるパラメータ推定，に分けられる．部分分布生存時間 RuleFit 法では，(a) に対して 2 種類の接近法が考えれる．一つは，Boosting tree (Friedman, 2001) に基づく接近法であり，もう一つは RandomForest(Breiman, 2001) に基づく方法である．このとき，後者では過度に大きな樹木を生成するため，Friedman & Popescu(2008) の流儀に倣い終結ふし数を規定する．(b) に対して，Ha *et al.*(2014) は lasso 法及び SCAD 法を部分分布比例ハザード・モデルに拡張している．(a) で得られたアンサンブル樹木は，(b) の過程においてダミー変数化されることを考慮すると，これらの縮小推定法が応用できる．本報告では，これらの組み合わせでの結果をシミュレーション及び文献例で評価する．