予測精度を用いたランダムフォレストによる生存解析について

東京理科大学　下川 朝有
東京理科大学　宮岡 悦良

はじめに

本研究では，生存時間を対象としたランダムフォレストの構築について扱う．ランダムフォレストの構築は，ブートストラップサンプルを用いることで複数の生存木を構築し，そのアンサンブルにより行われる．生存木の構築における最も重要な設定の一つは，各分割に対する評価基準があげられ，一般にはログランク検定等が用いられる．ログランク検定を用いる分割は，分割により得られる２つの集団に対して推定される生存時間関数間の距離を最大とすることに等しい．しかしながら得られるモデルの予測精度に着目した場合，C-index 等に代表されるモデルの予測精度に着目した基準を分割基準として用いることが考えられる．そこで本研究では，モデルの予測精度に着目した分割基準を用いたランダムフォレストの構築について考察を行う．

記述・一致確率

被験者 i の従う真の生存時間を表す確率変数を U_i，打ち切り時間を表す確率変数を D_i とし，その観測時間を $X_i = \min(U_i, D_i)$，イベント指標を $\Delta_i = I(X_i = U_i)$ とする．$\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \cdots, Z_{ip})$ を p 次元の共変量ベクトルとし，その取り得る値の集合（共変量空間）を $\mathcal{Z}$ とする．ここである関数 $h : \mathcal{Z} \to R$ により，被験者 i に対するイベント発生のリスクを表すスコア $H_i = h(\boldsymbol{Z}_i)$ が得られるとすると，観測値の集合は以下で与えられる：

$$\mathcal{L} = \{(x_i, \delta_i, \eta_i) : i = 1, 2, \cdots, n\},$$

ただし x_i, δ_i, η_i はそれぞれ，X_i, Δ_i, H_i の実現値を表し，リスクスコア H_i の値が高くなることは，イベント発生のリスクが高くなることを表すと仮定する．

ある時点 u における被験者 i の条件付き生存確率を $S_i(u) = Pr(U_i > u | H_i = \eta_i)$ とすると，その推定量 $\hat{S}_i(u)$ を用いることで，一致 (Concordant) は以下で定義される：

$$\begin{cases} U_i < U_j \text{ and } \hat{S}_i(u) < \hat{S}_j(u) \\ U_i > U_j \text{ and } \hat{S}_i(u) > \hat{S}_j(u) \end{cases} \Leftrightarrow \begin{cases} U_i < U_j \text{ and } \eta_i > \eta_j \\ U_i > U_j \text{ and } \eta_i < \eta_j \end{cases}$$

($j = 1, 2, \cdots, n$, $j \neq i$)．これを用いることで，与えられた生存モデル（リスクスコア）に対する観測内の一致確率 (Concordant probability) は以下で与えられる：

$$C = Pr(Concordant) = Pr(\eta_i > \eta_j | U_i < U_j, U_i < \tau),$$

ただし τ は C を評価する上での最大時間を表し，本研究では $\tau = \max\{x_i : \delta_i = 1, i = 1, 2, \cdots, n\}$ とする．実際は観測内に打ち切りが含まれるため，C に対する推定量を用いて評価する必要があり，その手法は幾つか提案されている．

木構造モデル

ランダムフォレストは複数の木構造モデル（生存木）のアンサンブルから成り，各生存木は共変量空間の分割ルールと，その結果得られる空間の部分集合（ノード）から成る．木構造を T，ノードを t で表すとすると，ノード t に対する分割ルールは “$\boldsymbol{Z} \in t_L$?” の形で与えられる．ここで $t_L \subset \mathcal{Z}$ 及び，$t_R = t - t_L$ は t の子ノードと呼ばれる．

各ノードにおける分割ルールを決定するためには，与えられた分割ルールに対する評価を求める必要があり，本研究ではこの評価に一致確率の推定値を用いる．分割ルールの評価を与えるため，ノード t_L に含まれる観測値は t_R に含まれる観測値よりも高いリスクを持つと仮定し，また同じノード内に含まれる観測値に対するリスクは等しいものと考える．すなわち同じノード内に含まれる観測値 i, j について，そのリスクスコアの値は $\eta_i = \eta_j$ であるとし，その組み合わせ (i, j) に対する C の推定値は 0.5 であると仮定する．