まえがき

コンピュータを基盤におく情報処理環境および計算環境が整備されるに伴い，ゲノム・データなどを含む大規模なデータ処理が短時間で容易に行えるようになりつつある．このような時勢のなかで，錯綜するデータに内在する構造（情報）を摘出するための方法が，統計科学あるいは情報科学（機械学習およびデータ・マイニング）の分野で研究開発されている．これらの分野において，活発に研究されている代表的な方法の一つが樹木構造接近法（tree structured analysis, tree based method あるいは recursive pertition と呼ばれる）およびアンサンブル学習法である．

樹木構造接近法は，応答に対する何らかの基準で応答に関与する説明変数を再帰的に分割しながらモデルを構築する統計的方法である．このとき，各反復における分割点の候補には，説明変数に含まれる観測値が用いられる．これにより，データに潜む要因構造が「樹木」によって視覚的に表現される．そのため，複雑な影響要因（変数）の非線形効果に対する鋭い洞察を与えるだけでなく，説明変数間の交互作用効果が自動的に摘出できる．このことは，「If ∼ Then」によるプロフィール（プロダクション・ルール）の作成に繋がる．

樹木構造接近法の出発点は，Mogan & Sonquist[106] の AID (Automatic Interaction Detection) 法である．ただし，AID 法は，数量化 I 類に対する交互作用の検出に焦点があてられており汎用性に乏しく，樹木構造接近法の特徴の一つである分割点を当初から固定して与えるために，分割点が解析の目標にそぐわないこともあり，その意味で広範に活用されなかった．

樹木構造接近法の用途を分類および回帰問題に拡張したのが，Breiman *et al.*[19] による CART 法 (Classification And Regression Trees) である．CART 法は，現在，多くの統計パッケージに実装されており，医学，環境科学，生態学，計量経済学，認知心理学といったさまざまな分野において応用例が報告されている．そのため，CART 法は，樹木構造接近法の「代名詞」になっている．CART 法では，ふし（ノード）内の不均一性測度（回帰問題の場合には残差平方和）の減少量が最大になるように，説明変数空間を再帰的に分割することでモデルが構築（推定）される．CART 法の提案以降，諸種の不均一性測度が提案されている．たとえば，一般化線形モデルの枠組みで導出された偏分残差に基づくふし内不均一性測度は，多変量応答，順序カテゴリカル応答あるいは生存時間解析といった諸種の回帰問題への拡張を可能にしている．とくに，生存時間研究における CART 法に基づく諸種の接近法の開発と応用が活発に行われている．これらの方法は，Survival CART 法と呼ばれ多くの提案と報告がある（たとえば，Davis & Anderson[34], Gordon & Olshen[61], Kehl & Ulm[91], LeBranc & Clowley [98] など）．がん臨床試験にお

ける事後の探索的な解析への有用性が SWOG (South West Oncology Group) の統計家である Crowley *et al.*[32] あるいは Green *et al.*[66] によって主張されている．

CART 法では，分割されたふし内の不均一性測度の減少量を基準にしている．他方，分割された娘ふし間での応答の相違を検定統計量および（多重比較調整を伴う）p 値で評価する方法もある．このような方法として有名なのが Kass[89] によって提案された CHAID (Chi-squared Automatic Interaction Detection) 法である．CHAID 法は，IBM 社のソフトウェア SPSS Decision Trees に実装されており，社会科学における計量的分析において広く活用されている．

CART 法および CHAID 法における応答の予測値は，終結ふし（リーフ）内の代表値（回帰問題の場合には，ふし内の平均値）によって与えられる．したがって，回帰問題におけるこれらの方法による潜在的な（真の）モデルに対する近似は，ステップ関数に基づいている．そのため，潜在的なモデルが線形構造をもつとき，推定モデルの近似確度が極端に低いか，あるいは多くの終結ふしを伴うモデルが構築される．

このような難点を解消するために，Friedman[49] は，ステップ関数近似を打ち切りベキ乗基底関数近似に変更し，局所的に線形なモデルをあてはめる方法，すなわち，MARS (Multivariate Adaptive Regression Spline) 法を提案している．実地において，データに内在する潜在的なモデルが全体的に複雑であるとしても局所的には線形な構造を示し，そして少数の説明変数しか関与しないかもしれない．MARS 法ではこのことに着目し，上記の方法と同様に，説明変数の空間を局所的な（説明）変数効果を最大限に引き出せるように再帰的に分割している．また，その分割された部分空間に区分線形関数をあてはめることで，連続な回帰モデルを構築している．再帰的なモデル構築の戦略は，非線形回帰モデルにおける説明変数間の交互作用効果を自動的に摘出することにも繋がる．

樹木構造接近法が諸種の分野において実践されていくにつれて，新たなニーズも増加している．たとえば，医学・医療の分野では，任意の治療法（あるいは薬剤）に対する適応患者像（レスポンダー）の摘出および患者像の探索が研究主題の一つである．このような場面では，応答を予測するというよりも，むしろ関心のある応答をもつ部分集合の抽出と，説明変数に基づく部分集合のプロフィール（プロダクション・ルール）に関心がある．このような場面に有用な方法がデータ・ピーリング（バンプ・ハンティング）法である (Friedman & Fisher [50])．データ・ピーリング法とは，関心のある応答をもつ部分集合を，説明変数の座標軸に沿って再帰的に探索する方法である．データ・ピーリング法の再帰アルゴリズムは，CART 法での戦略の類推により構成される．

近年，諸種のアンサンブル学習法が提案され，任意の弱い機械学習器を強力にすることに成功している．CART 法は諸種の長所を保持するが，何らかの線形モデルと比べてもそれほど予測確度が高くないことが弱点である．MARS 法は，CART 法を連続モデルに拡張することでより高い予測確度をもつ．しかしながら，MARS 法は CART 法のように説明（予測）変数の単調変換に不変ではない．このとき，CART 法の長所を保持しながらアンサンブル学習法を介して，より強力な予測確度をもつように組み立てられた接近法が，MART (Multiple Additive Regression Trees, Friedman [51]) 法，Bagging 法 (Breiman[12])，あるいは RandomForest 法 (Breiman[16]) である．このとき，MART 法と Bagging 法および RandomForest 法では，異なったアンサンブル戦略がとられる．前者では，各樹木に対してあてはめる応答の値を逐次に更新するブースティング戦略 (Freund [46]) が用いられ，後者では，ブートストラップ標本に対して樹木をあてはめる戦略が用いられる．

AID 法から出発した樹木構造接近法は，CART 法の提案により用途が拡大し，CART 法におけるモデル構築のアルゴリズムを踏襲する形式で諸種の方法が提案されている．さらに，コンピュータの高速化によって，より精緻化された統計的モデルの構築が可能になってきている．このように樹木構造接近法の研究と開発が目覚ましく発展していくなかにあって，それらの方法を包括して解説した著作は出版されていない（Hastie *et al.*[73] のなかで樹木構造接近法の多くの手法が取り上げられているが，この成書は統計的機械学習法の総合的な解説書であり，樹木構造接近法の周辺的な内容には触れていない）．また，和文に至っては，樹木構造接近法および周辺諸法を詳細に解説した著作は，皆無である．

そのため本書では，樹木構造接近法に関する著者らの総合報告 [145] に基づいて，個々の手法に関する具体的内容を取り扱う．そこでは，樹木構造接近法およびアンサンブル学習法の理論的な背景，および周辺手法に関する話題も取り上げる．紙面の都合で十分に解説できない場合には，関連する多くの文献を紹介することで不十分な内容を補完する．

さらに本書では，諸種の樹木構造接近法，アンサンブル学習法およびその周辺諸法に関する理論的内容だけではなく，統計ソフトウェア R を用いた統計的データ解析の実践を取り扱う．本書で解説されるほとんどの方法は，R のパッケージとして公開されており（2013 年 7 月現在），インターネット環境が整備されているコンピュータ上で再確認が可能である．本書で使用するパッケージの詳細な内容は CRAN (`http://cran.r-project.org/`) で見ることができるが，英語表記であることからとっつきの悪い印象を与える．本書では，手法の実行に最低限必要な引数に関する略説を記している．

R は，統計ソフトウェアのデファクト・スタンダードになりつつあるものの，一方で，不十分な出力結果しか得られないパッケージおよび関数が散見される．本書では，このような出力に対して，新たな関数を作成することで統計的データ解析に対する環境の整備を図っている．また，適用するデータについても，R のパッケージのなかに含まれているものを用いることで，データのダウンロードなどの煩わしさを解決している．

本書の構成を以下に示す．第 1 章では，CART 法およびその拡張の諸型について解説する．R では，`rpart` および `mvpart` の 2 種類のパッケージが存在する．後者は前者の拡張パッケージ（ラッパー）として開発されているものの，そのデフォルト値あるいは交差確認法の有無など，いくつかの点に違いがある．R における CART 法の応用については，本シリーズを含めたいくつかの文献で見ることができる．しかしながら，これらのパッケージの違いについて述べられたものはない．そのため本書では，それぞれの方法の適用方法および留意点について詳細に述べる．さらに，Poisson 回帰樹木，および多変量回帰樹木といった，諸種の応答に対する CART 法の背景と R による実行方法について解説する．

第 2 章では，検定統計量に基づく樹木について記す．そこでは，条件付き推測樹木 (Hothorn *et al.* [79]) について解説する．近年，データ・マイニングの分野において，樹木モデルと線形モデルの混合形式で構成される手法が提案されている．これらの手法は，ハイブリッド型樹木と呼ばれており，樹木モデルの構築には，検定統計量に基づく方法が採用されている．そのため，第 3 章において，これらの手法の概説および R での実行手順について触れる．

第 3 章では，MARS 法とその周辺について述べる．R における MARS 法のパッケージには，`mda` および `earth` が存在する．本書では，これらのパッケージの違いと適用上の留意点について

詳説する．さらに，MARS 法の周辺諸法として，柔軟判別分析 (Hastie *et al.*[71]) および論理回帰法 (Ruczinski *et al.*[116]) を概説し，R による実行方法について述べる．

第 4 章では，データ・ピーリング法として，PRIM (Patient Rule Induction, Friedman & Fisher [50]) 法について述べる．R における PRIM 法のパッケージには，`prim` があるものの，その出力および診断グラフィクスは整備されていない．そのため，本書ではこれらを補完するための関数を構成する．データ・ピーリング法は，アソシエーション・ルール分析と類似しているように思われがちであるが，異なる目標および戦略により構成される．前者は，外的基準がある場合のデータ解析手法であり，CART 法のアルゴリズムを踏襲している．他方，後者は，外的基準をあらかじめ設定せずに，連関 (association) の高いアイテム集合を抽出する．また，連関の高いアソシエーション・ルールの抽出には，条件付き確率の推定値を用いる．これらの違いを表すために，本章では，アソシエーション・ルールの概要を統計的な視点から述べ，そして，R による実行の方法について説明する．

第 5 章では，ブースティング戦略に基づくアンサンブル学習法について触れる．ブースティング戦略の原型は，Tukey の成書「探索的データ解析 (Exploratory Data Analysis) [123]」のなかの twicing に見ることができる．ただし，現在の発展の背景には，Freund & Schapire による AdaBoost 法 [47] の提案が深く関与している．そのため，本章では，AdaBoost 法の概説から出発し，その発展形である MART 法について述べる．また，本章では，第 6 章のブートストラップ法戦略に基づく方法との違いについても触れる．

第 6 章では, ブートストラップ戦略に基づくアンサンブル学習法について触れる．そこでは, まず, Bagging 法について述べ，次いで RandomForest 法について触れる．このとき，RandomForest 法については，二三のグラフィカル診断の方法について議論する．

本書では，それぞれの章を独立して学習するために，R のコマンドが重複している個所がある．これは，樹木構造接近法，アンサンブル学習法に対する諸法のリファレンスとして活用できることに留意しているためである．R は，MacOS，Linux および Windows といったマルチプラットホームに対応しているが，本書はすべて Windows で実行している．また，R のバージョンは，3.0.1 である．さらに，本書で取り上げたパッケージおよび関数は，すべて 2013 年 7 月現在のものである．

本書は，樹木構造接近法およびアンサンブル学習法に焦点をあてた，本邦における初めての成書である．本書が読者の研究および学習に対して少しでも寄与することを祈念し，まえがきの結びに代える．

2013 年 7 月

著者を代表して　下川 敏雄

謝　辞

本書の出版にあたり，金 明哲先生（同志社大学）には，執筆を勧めていただいただけでなく，本書の構成などでたいへんお世話になった．心から御礼申し上げる．本書のためにさまざまな助言をいただいた NPO 法人 医学統計研究会の諸先生方には，謝意を表したい．共立出版（株）の横田穂波氏には，本書の校正だけでなく，著者らの執筆が予定を大幅に遅れ，ご心配をおかけした．ここに，お礼とお詫びを申し上げる．本書に至るまでに，各著者が，それぞれの環境で，多くの方々にお世話になった．以下に，各著者からの謝辞を記したい．

樹木構造接近法の研究にあたり，北村眞一先生（山梨大学）ならびに古河 洋先生（近畿大学）には，多くの実践場面を提供いただいた．さらに，辻 光宏先生（関西大学）および坂本 亘先生（岡山大学）には，本書の執筆に関してさまざまな助言をいただいた．また，後藤昌司先生，松原義弘先生（医学統計研究会）ならびに医学統計研究会の先生方の「研究財産」は，本書の構成において多大な影響を与えてくれた．さらに，杉本知之先生には，本書の動機となった総合報告 [145] の執筆だけでなく，本書においても丁寧な推敲と修正をいただいた．謝辞には書ききれないほどの諸先生方や学生たちのさまざまなご支援が，本書の制作において多大に寄与している．この場を借りて，心からお礼申し上げる．最後に，研究室に籠って本書を執筆する姿を間近で見守り，心配をかけた妻 こずえに心から感謝したい．　下川 敏雄

総合報告 [145] では，構想から脱稿までに数年を要した．これは主に自身の怠慢によるが，樹木構造接近法の深さと進展の勢いにも一因があった．同時に，Breiman や Friedman らの天才的閃きに触れ，時折難解な彼らの著作の理解のため，Salford Systems 社のソフトと自作 Fortran の結果とを一つずつ突き合わせる必要があった．本書は，[145] の割愛部に加え，下川敏雄氏の主たる労力のもとで新たに加えられた方法と R の事例を含む．本書に至るまでに多くの知識の提供を頂いた濱崎俊光先生（大阪大学），池田公俊氏，古川泰伸氏に感謝したい．　杉本 知之

樹木構造接近法の研究と開発は，著者がシオノギ製薬（株）の解析センター長として精励していた 1984 年頃から始まった．仲間とともに多くの統計的データ解析における接近法を日本で最初を競って手がけ，研究・開発してきた．CART 法も日本で最初に FORTRAN で仲間とともに

開発し，MARS 法についても Friedman (1991) のぶ厚い論文を討論部分も入れて仲間内で全訳・整理して開発した．1995 年に大阪大学に移ってからは，若い仲間とこれらの方法を発展させた．とくに，生存時間解析では，杉本知之君が先頭に立ってくれて多くの方法を開発してくれた．また，本書の構成を率先して進めてくれた下川敏雄君は，計算環境や情報環境が著しく速く進む中にあって，その能力を発揮し，これまでの成果をまとめてくれた．ここに，シオノギ製薬（株）解析センターでの仲間の方々，とくに松原義弘（博士），勘場 貢，山本成志，惣田隆生の方々に，心より謝意を表したい．また，生存時間研究での樹木構造接近法の実践の中で，ご指導とご討議頂いた中島聰總先生，古河 洋先生，前谷俊三先生にこの場を借りてお礼申し上げる．　　後藤昌司

記　法

記号	定義・例	説明
一般		
n	$n = 1, 2, \ldots, N$	標本番号を表す添え字
p	$p = 1, 2, \ldots, P$	説明変数番号を表す添え字
k	$k = 1, 2, \ldots, K$	クラス番号を表す添え字
$\mathscr{L}$	$\mathscr{L} = \{y_n, \boldsymbol{x}_n\}_{n=1}^{N}$	データ集合
$\boldsymbol{X}$	$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)^{\mathrm{T}}$	P 個の説明変数の N 個の標本からなる $N \times P$ 行列
$\boldsymbol{x}_n$	$\boldsymbol{x}_n = (x_{1n}, x_{2n}, \ldots, x_{Pn})^{\mathrm{T}}$	P 個の説明変数からなる n 番目の標本
$\boldsymbol{Y}$	$\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_Q)$	Q 個の応答の N 個の標本からなる $N \times Q$ 行列
$\boldsymbol{y}$	$\boldsymbol{y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$	単一応答の N 個の標本からなる $N \times 1$ ベクトル
$\widehat{\boldsymbol{y}}$	$\widehat{\boldsymbol{y}} = (\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_N)^{\mathrm{T}}$	応答 $\boldsymbol{y}$ の推定値
$\overline{\boldsymbol{x}}$	$\overline{\boldsymbol{x}} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_P)^{\mathrm{T}}$	P 個の説明変数の $P \times 1$ 平均ベクトル
$\boldsymbol{\beta}$	$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^{\mathrm{T}}$	P 次元の回帰（判別）パラメータの $P \times 1$ ベクトル
$\mathrm{N}(\cdot, \cdot)$	$\mathrm{N}(\mu, \sigma^2)$	正規分布
$\mathrm{MVN}(\cdot, \cdot)$	$\mathrm{MVN}(\boldsymbol{\mu}, \Sigma)$	多変量正規分布
Υ		未知の確率分布
$\boldsymbol{\theta}$	$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_P)^{\mathrm{T}}$	パラメータ
μ		平均（位置）パラメータ
σ^2		分散（尺度）パラメータ
$\boldsymbol{\Sigma}$		分散共分散行列
$\mathrm{E}[\cdot]$	$\mathrm{E}[Y]$	期待値
$\mathrm{Var}[\cdot]$	$\mathrm{Var}[Y]$	分散
pr	$pr(j\|k)$	確率
$\mathbb{1}(\cdot)$	$\mathbb{1}(x \leq c)$	括弧内が正ならば 1，負ならば 0 を返す指標関数
$\mathrm{sign}(\cdot)$	$\mathrm{sign}(x \leq c)$	括弧内が正ならば 1，負ならば -1 を返す符号関数
V	$L_{\mathrm{sub}}^{v}, v = 1, 2, \ldots, V$	交差確認法における部分標本の数
p		p 値
α		有意水準
一般（樹木）		
l	$l = 1, 2, \ldots, l_e$	ふし番号を表す添え字
j	$j = 1, 2, \ldots, J$	樹木系列を表す添え字
o	$o = 1, 2, \ldots, O$	分岐点候補番号を表す添え字
b	$b = 1, \ldots, B$	アンサンブル回数を表す添え字
t_l	$t_1 \in \{1, \ldots, N\}$	l 番目のふし（ノード）に属する個体集合
$\|t_l\|$		ふし t に属する個体数
$N(t)$		ふし t に属する個体数
T	$T = \cup_{l=1}^{l_e} t_l$	すべてのふしの集合
$\widetilde{T}$	$\widetilde{T} = \{t_4, t_5, t_7, t_8, t_9\}$	終結ふしの集合
$\|\widetilde{T}\|$		終結ふしの個数
$p(t)$		ふし t における分岐変数
$c(t)$		ふし t における分岐点
Imp_p		変数 x_p の変数重要度
$h(\cdot)$		基本学習器