決定木学習：実用期を迎えたノンパラメトリック統計手法

同志社大学大学院　金田 重郎 [8]
skaneda@mail.doshisha.ac.jp

決定木は、対象物の属性について質問を積み重ねてゆくことにより、最終的にある種の判断を実現する知識表現である。専門家の判断事例を蓄積し、この事例から自動的に決定木を生成する決定木学習手法の発達により、近年、SPSS等、多くの統計解析ツールにパッケージとして組み込まれるようになった。決定木は、線形モデルを用いた従来の統計手法とは異なり、非線形の知識表現であり、知識記述能力が高い。本稿では、決定木学習の代表的な手法であるID3,CARTについて紹介して、決定木学習の特長と適用範囲について紹介する。

1　はじめに

決定木モデル (樹形モデル) は、対象物の属性について質問を積み重ねてゆくことにより、最終的にある種の判断を実現する知識の表現方法である。古くから使われていた表現であるが、近年、ID3,CART等の、事例から自動的に決定木を導く手法 (機械学習アルゴリズム) が種々開発されたことにより、顧客データ分析等の多くの応用への適用が活発化している。

決定木は、従来の重回帰分析等のパラメトリックな統計手法とは異なり、事前の事例分布や誤差に何らの分布も仮定しない、ノンパラメトリックな統計手法である。有力な統計解析パッケージであるS-PLUSには以前から搭載されていた[1]が、数学的基盤が弱いと見られ、統計利用分野では、さほど重要視されてはいなかった。

これに対して、人工知能 (AI) 分野では、機械学習の重要な1分野として、この10年程の間、盛んに研究され、実用的な手法が種々開発されている。代表的手法にID3がある。

ポイントカード、電子商取引 (EC) の進展により、大量のデータが蓄積され初めている状況から、最近では、決定木は、SPSS等、多くの統計解析ツールに組み込まれている。線形モデルを用いた従来の統計手法とは異なり、非線形の知識表現であり、知識記述能力が高い。

本稿では、ID3に焦点を絞りながら、決定木学習手法の原理・利点について紹介したい。

[1]S-PLUSでは、樹形モデルと呼んでいる。

2　機械学習と決定木

2.1　機械学習

機械学習とは、事例からその背後にある一般的法則を帰納する技術を言う。決定木学習はその一分野であり、現実の社会に存在する事例 s_j を以下の形で表現する。

$$s_j = (a_{1,j}, a_{2,j} \cdots a_{i,j} \cdots a_{n-1,j}, a_{n,j} : C_j) \quad (1)$$

ここで、j は事例番号であり、事例の個数を M とするとき、$j = 1, 2 \cdots M$ である。$a_{i,j}$ は属性値である。**属性**とは、例えば、判断すべき対象が病名であり、事例が個々の患者であるとしたとき、患者個々の性別、咳の有無、発熱の有無、体重等が属性であり、その値を**属性値**と呼ぶ。ここでは、属性値に、1から n までの番号を付した。n は、属性の個数である。C_j は事例 s_j が有するクラスであり、例えば、患者 s_j に対する病名である。これら属性からクラスを決定する知識を自動的に獲得することが、決定木学習の任務となる。即ち、病気の診断であれば、病名未知の患者を診断する知識を決定木として診断事例から得ることになる[2]。

機械学習とは、**M 個の事例（学習事例）が与えられた時に、クラスを推定する知識を自動的に生成する**ことである。但し、少ない事例からでも、高精度の知識が得られることが望ましい。

[2]統計分野では、クラスを被説明変数、属性を説明変数と呼ぶ。この呼び名の違いは習慣的なものであるが、属性値が数値ではないことを強調するために、本稿では、「属性」を利用する。

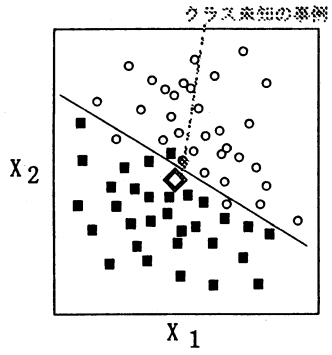

図 1: 線形分離の例

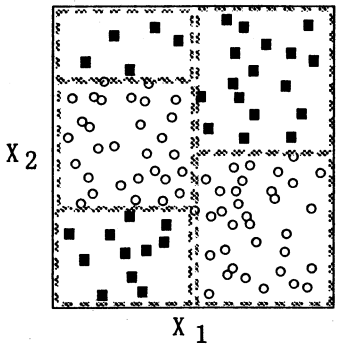

図 2: 線形非分離の例

2.2 線形分離

図 1 は、2 個の属性,x_1,x_2 を持つ事例を 2 次元的に配置したものである。事例の持つクラスを○と■で表現している。例えば、過去の患者が、その症状により配置され[3]、事例対応のクラスである○または■が 2 種類の病名に相当する。

前述の目標「クラスが未知の事例に対して、高い精度でそのクラスを推定する知識を得る」について考えてみる。例えば、図 1 中の◇のように、属性値は分かっているが、クラスが未知の事例が得られたとする。病気の症状は問診できる。◇が○であるか■であるかを判断する必要がある。この場合、図 1 に示したように、領域を 2 分する直線[4]をあらかじめ引いておけば、直線のどちら側に入るかで、○か■かを判断できる。図 1 では、◇は■と推定できる。この直線による領域分割が、「高い精度でそのクラスを推定する知識」である。

図 1 のように、空間を直線で分割するものを線形モデルと呼ぶ。重回帰分析や判別分析等の統計解析手法は、基本的には線形モデルを用いている。数学的に扱いやすく、図 1 の様に、線形分離できない事例があったとしても、誤差を最小化すること (残差平方和の最小化) を線形代数の範囲で保証できる。

2.3 線形非分離

しかし、直線による分割 (線形モデル) には、大きな問題点がある。図 2 のような場合はどうしたらよいのであろうか？この例では、1 本の直線で領域を分割することはできない。この場合 (線形非分離) には、図 2 で示すように、複数の領域に分割することが望ましい。

一方、統計分野では、「**オッカムの剃刀**」と呼ばれる大原則があり、「クラスを推定する知識」は、できるだけシンプルであるほうが、将来の事例に対する推定性能が良い。従って、機械学習の課題は以下に集約される。

【目標】:多次元空間上に分布する事例 (これを**学習事例**と呼ぶ) を、最小の個数の多次元直方体 (これを以下、キューブと呼ぶ) でカバーし、キューブ毎にクラスを 1 個定める。キューブ内の事例と、キューブのクラスは一致しているものとする。

上記目標に対して、以下のことが分っている[5]。

【効率的計算手法の非存在:】最小個数となるキューブを求める「効率的」方法は存在しない。ここで、効率的とは、事例数を M とするとき、カバーを求める時間が、M の 2 乗や 3 乗といった指数関係を遙かに越えることを言う。現実的なサイズの問題では、解を見つけることは不可能である。

即ち、図 2 のような、最小個数のキューブによるカバーを求めることはできない。であるからこそ、従来の統計解析では、空間分割能力が劣ることを承知の上で、線形モデルに限定し、その代償として、数学的厳密性を得ていたのである。

2.4 決定木とは何か？

以上のような背景から、非線形のモデルを用いた推定手法として、研究されてきたのが、決定木

[3]実際の病名の診断では、属性は多数あるので、多次元空間上の点となるが、ここでは、説明のために、2 次元とした。

[4]多次元空間上の超平面

[5]現実には、属性値に誤り等があり、図 1 のように、他キューブに入ってしまう学習事例 (**ノイズ事例**) が存在することが多い。ノイズ事例も含めて全事例を厳密に分類すると、キューブ数が多くなり、未知事例への推計性能が低下する。ここでは、ノイズ事例はないとしておく。

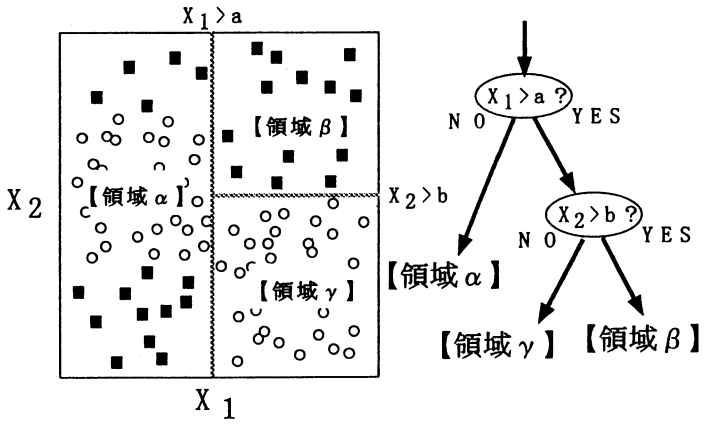

図 3: 決定木の原理

である。図 3 を用いて、その考え方を示す。

決定木では、「分割統治 (Divide and Conqure)」により、判別知識を生成する。即ち、学習事例が分布する空間に対して、まず、全体を眺めて、ある属性により大きく分ける。図 3 の例でみると、まず、属性 X_1 に着目して、それが a よりも大きいか小さいかで、全体の領域を左右に 2 分割する[6]。

次に分割後の右側領域に注目する。上半分と下半分で、クラスが分かれている。そこで、中央付近 ($X_2 = b$) で、2 領域に分ける。これにより、少なくとも、右半分の領域では、学習事例のクラスは、きれいにそろう。これらの領域 (キューブ) に未知事例が飛び込んで来た場合には、その未知事例のクラスは、右の上半分は、「■」、下半分は、「○」として良い。図 3 の左半分については分割を続ける必要がある。

以上をグラフ的に表現したものが、図 3 右側の決定木である。分割統治で領域を分割してゆき、分割された領域中のクラスが一意になった時点で、その枝については、分割を止める。枝別れの部分を「**ノード**」とよび、クラスが一意となってキューブが完成したものを、「**リーフ**」と呼ぶ。このような構造全体を「**ツリー (木)**」と呼ぶ。

具体的には、領域 β については「■」、領域 γ については「○」とする。左側の枝についても、同様の処理をすることにより、決定木が完成する。キューブが求まれば、学習事例はすてる。図 3 の右側に示したような決定木のみが保存され、今後生じてくる事例のクラス判定に供される。

新しい (クラスが未知の) 事例が到着した場合には、この決定木の上から、その事例を流して、属性値に対するテストを行う。例えば、図 3 において、新しいクラス未知の事例 F を考える。この事例 F では、X_1 の値は a より大きな A であり、X_2 の値も b より大きな B であったとする。決定木の最上部から、事例 F を流す。最初に、X_1 について、テストが行われ、その結果 F は、右の枝に流れてゆく。次に、X_2 がテストされ、更に右下の枝へと流れてゆく。このノード (領域 β) では、学習事例から、すでにクラスが「■」であるとされているので、この未知事例 F のクラスは「■」となる。

2.5 決定木学習アルゴリズム

複数の属性と、唯一のクラスを持つ事例を考え、その事例の集合を T とする。T は**学習事例**と呼ばれる。この時、事例の持つクラスは、$\{C_1, C_2, C_3 \cdots C_j \cdots C_n\}$ とする。n は、クラスの個数である。決定木作成のプロセスは以下のようなものである。

*STEP*1 : T がすべて同一のクラス C_j であるならば、T に対する決定木はリーフであり、そのクラスは C_j である。

*STEP*2 : T が全く事例を含まない空である場合には、上記 *STEP*1 と同様にリーフとするが、そのクラスは、事例からは決めれない[7]。

*STEP*3 : T が複数のクラスを有する場合には、ある属性を選択して、その属性が持ちうる属性値により、事例 T を分割する。そして、

[6] どうやって、分割に利用する属性と属性値とを定めるかについては、後述する

[7] この決め方は、ツールに依存する。後述の C4.5 では、ひとつ上流側のノードへと流れて来た事例 T のクラスのなかで、最頻クラス。

(高い、黒色、茶：-)
(高い、黒色、青：-)
(低い、黒色、青：-)
(高い、ブロンド、茶：-)
(低い、ブロンド、茶：-)
(低い、ブロンド、青：+)
(高い、赤色、青：+)
(高い、ブロンド、青：+)

図 4: 学習用データ

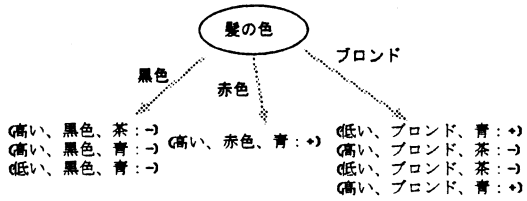

図 5: 髪の毛による分割

分割された事例集合の各々について、上記 $STEP1 \sim STEP3$ を再帰的に実行する。

3 決定木学習アルゴリズム ID3

上記の決定木学習で、分岐させる属性の決定方法が問題となる。生成された決定木の性能に大きな影響を与える。代表的な決定木作成アルゴリズム ID3[3] のエントロピーを利用した属性選択法について紹介する。

まず、学習事例として、図 4 に示す 8 個を考える。各事例は、「背の高さ」「髪の毛の色」「目の色」の 3 個の属性と、2 値のクラスから構成される。属性値は、背の高さが 2 通り、髪の色が 3 通り、目の色が 2 通りなので、考え得る事例の種類は $2 \times 3 \times 2 = 12$ 通りであるので、4 種類の事例は未収集である。未知事例を含めた、全事例のクラスを判定する決定木を生成する。

Quinlan の ID3 では、上記のような分割の良さを評価する手段として、情報エントロピーを用いる。情報エントロピーは、情報の乱雑性を評価するために用いる量であり、クラス数が 2 の場合には、「+」「-」のクラスの発生確率を p_+, p_- とする時、以下の式で与えられる。但し、log は log_2 であり、エントロピーの単位は bit である。

$$E = -p_+ \cdot log(p_+) - p_- \cdot log(p_-) \qquad (2)$$

従って、図 4 のエントロピーは

$$E = -(\frac{3}{8}) \cdot log(\frac{3}{8}) - (\frac{5}{8}) \cdot log(\frac{5}{8}) = 0.954bit \quad (3)$$

となる。

次に、髪の色で、この事例を分けてみる。図 5 のように、事例は 3 個の集団に分離する。ここで、髪の色が黒と赤の場合には、クラスがそろっており、前述の $STEP1$ から、これがリーフとなる。一方、髪の色がブロンドの場合には、クラスが分かれており、引き続き、髪の色以外の属性での、分割を繰り返す必要がある。

髪の毛による分割以前と以後で、エントロピーはどう変化するかを考える。式 (2) の定義から、この場合、髪の色が黒、または赤の場合にはエントロピーはゼロである。エントロピーゼロとは、まったく乱雑性がない状態である。ブロンドの場合には、「+」と「-」が半々なので、エントロピーは、以下のようになる。

$$E = -(\frac{2}{4}) \cdot log(\frac{2}{4}) - (\frac{2}{4}) \cdot log(\frac{2}{4}) = 1bit \quad (4)$$

そして、分割後のエントロピーの期待値は、それぞれの枝に流れている事例数を考慮して、以下のように求める。

$$(\frac{3}{8}) \cdot 0 + (\frac{1}{8}) \cdot 0 + (\frac{4}{8}) \cdot 1 = 0.5bit \qquad (5)$$

従って、分割前 (式 3) と分割後 (式 5) のエントロピーの差は、

$$0.954 - 0.5 = 0.454 \qquad (6)$$

となる。

以上から、「髪の色」を属性として選択した時のエントロピーのゲインは、0.454bit である。同様にして、他の属性 (背の高さ、目の色) のエントロピーゲインは、(1) 背の高さ：0.003 bit、(2) 髪の色:0.454bit、(3) 目の色：0.347bit、である。

髪の色による分割は、エントロピーゲインが一番大きい属性であったことが分かる。エントロピーゲインが大きい事は、事例集合の乱雑性が大きく減少することを意味する。即ち、図 2 の例のように、うまく事例をカバーする可能性が大きいことになる。このようにして、ID3 では、エントロピー削減が一番大きな属性をつぎつぎと再帰的に利用しながら、枝を伸ばして、決定木を構築する。

決定木の目標は、できるだけ少ないノード（属性値の質問）数で、決定木を構成することである。これは、言い換えると、図 2 のように、多次元空間をできるだけ少ないカバーで覆うことを意味している。最小ノード数の決定木を構成する効率的な方法は本質的に存在しない。しかし、情報エントロピーに基づく方法は、簡単な方法ながら、優れた近似解を導くことが知られている。

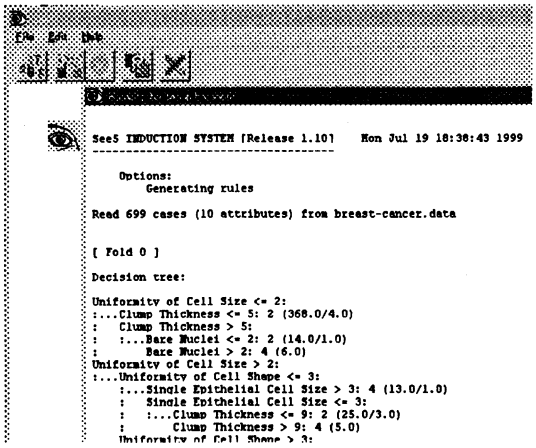

図 6: See5

但し、上記の原理的手法だけでは、現実の問題を扱うことは困難であり、種々の改良が加えられた。問題点と対処策を以下に示す。

過学習の防止：現実の問題では、測定誤差等により、別クラス領域に、事例が入り込むことがある。この場合、事例集合のクラスが一意になるまで枝の展開を繰り返すと、決定木が過度に大きくなり、推計性能が低下する。これを「**過学習**」と呼ぶ。この問題を解決するため、ID3 改良版の C4.5[4] では、決定木の成長を途中で止めたり、あるいは、一度成長した決定木を刈り込む機能を持っている。これを、「**プルーニング**」と呼ぶ。

過度に平坦な木の防止：上述のエントロピーに基づく属性選択は、「平たく横に幅広い木」を作りがちである。この問題を回避するため、C4.5 では、*split info* と呼ばれる評価尺度を利用している。

Missing Value への対処：現実の事例では、ある属性値が何らかの理由で測定されておらず、属性値が不明の場合がある。これを「**Missnig Value**」と呼ぶ。C4.5 では、この種の事例に対しても対処可能としている。

数値属性の許容：以上の説明では、属性値はあくまで飛び飛びのシンボル値であり、その順序には意味がない (名義尺度) とした。しかし、現実の応用では、数値属性がしばしば現れる。この様な属性に対して、C4.5 では、自動的にあるスレショールド (閾値) を求めて、そのスレショールドより大きいか否かを尋ねる質問を生成する。これにより、数値属性と名義尺度属性とが混ざった事例を許容する。

尚、C4.5 では、決定木形式のみではなく、ルール形式の知識も生成するようになっている。ID3 は、人工知能分野では極めて有名な手法であり、これを利用した研究や改良版が多い。その最大の特徴は、属性を原則として名義尺度としていること、そして、枝分かれが 3 以上の複数分岐 (マルチスプリット) である点である。後述の CART では、マルチスプリットは実現できず、名義尺度が多い応用は得意とは言い難い。

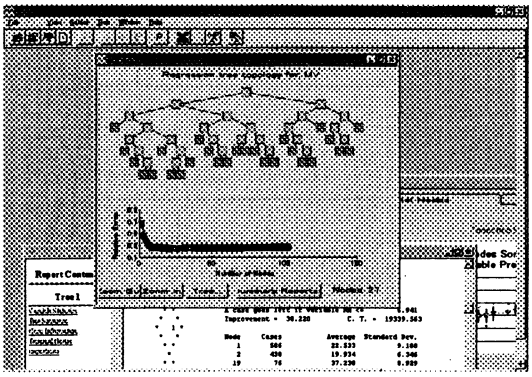

図 7: Salford 社 CART

4 2進木学習アルゴリズム CART

ID3 とは異なり、統計家に好感を持って迎えられている決定木学習アルゴリズムに CART がある。CART は、属性値 (場合によってはクラスも) が連続値であることを基本とする。分岐は 2 分岐に限定される。このため、「2 進木」の名前がある。CART では、クラスが名義尺度の場合には、「分類 2 進木」とよび、クラスが連続値の場合には、「回帰 2 進木」と呼ぶ。

CART も、分割統治により、領域を分割して、決定木を作成するところでは、ID3 と同一である。しかし、分岐すべき属性 (分岐変数) の選択には、エントロピーではなく、GINI インデックスを利用している。例えば、あるノード t において、事例集合は、C 個のクラスに分散しているとする。この場合、各クラス毎の事例が、事例全体の中に生起する確率を $p(j|t)$ で表すとき、GINI インデックスは以下の式で表される[8]。

$$GINI = 1 - \sum_{j}^{C} p(j|t)^2 \qquad (7)$$

[8]例えば、クラス 2 の場合について考えてみると、クラスが一方に偏っていれば、この GINI 分散指標も 0 となり、逆に等確率にバランスしていると最初値になる。このことから、この GINI インデックスとエントロピーの両者間には、大差はない。

ID3との最大の違いは、名義尺度の属性に対する振る舞いである。ID3では、属性値が3種類以上存在しても、決定木の展開に支障はない。しかし、CARTでは、もともと2分岐しか考えていない。このため、CARTでは、名義尺度の変数であることを陽に宣言すると、(属性値がn種類あるとして)n個を2グループに分割するすべての組み合わせを調べる。nが大きくなると、組み合わせは膨大となる。このことからも、CARTでは、従来の統計解析と同様に、属性としては、連続値（数値）を強く意識している。尚、CARTでもMissing Value、過学習に対する対策はID3と同様に盛り込まれている。

CARTのID3と比較した利点の一つは、クラスが連続値でもよいことである。これを、回帰2進木と呼ぶ。回帰2進木では、群間平方和を最大化(群内平方和を最小）することを評価基準として、木を展開するが詳細は文献[4]に譲る。但し、頑健性（ロバストネス）のある評価値として、2乗和ではなく、偏差(誤差の絶対値)を用いることも可能としている。また、CARTが用いているGINI分散指標では、2分岐した一方に流れる事例が極端に少ない選択が有利となる欠点が知られており、この問題に対処するため、Twoing Ruleと呼ばれる、改良された指標も準備されている。

5 決定木の適用分野

5.1 商用パッケージ

ID3及びCARTは、ソフトウエアパッケージとして市販されている。図6は、ID3の商用版であるRulequest社[6]のSee5の動作画面である。但し、C4.5は、Quinlanの著書[3]に付属しており、研究目的であれば無料で利用できる。C5.0及びSee5は、C4.5の機能強化版である。

一方、CARTは、Salford社[5]より市販されている。動作例を図7に示す。CARTでは、種々のデータにより決定木を分析できる。ソフトウエア製品としては、前述のSee5より完成度が高い[9]。

5.2 決定木の適用分野

決定木の適用範囲については、研究を待たなければならない点が多い。しかし、一般的には、以下が言われている。

可読性の高さ：名義尺度を利用でき、しかも、得られた知識が人間に可読である。従来の重回帰分析では、ダミー変数とよばれる「0」「1」の2値を取る属性を、各シンボル値に対して割り付けていたので、直感的理解には向かない。

部分構造抽出：属性Aがある値では属性Bが影響し、属性Aが別の値となるとBは無関係でCが影響する場合でも、決定木は知識を表現できる。これに対して、重回帰分析では、属性の部分集合を取り出すので、細かい部分構造を解析できない。

大きな決定木の出現：多数の属性の値を決めない限りクラスが決まらない場合については、注意が必要である。この種の問題を決定木記述すると、決定木のどの分岐も、すべての属性を通るまで枝を伸ばすので、決定木が巨大となる。

学習時間の短さ：決定木学習手法は、ニューラルネットのような他の学習手法に比べて、処理が高速である。このため、巨大データ解析や、ブートストラップ手法[7]に適している。

6 おわりに

決定木学習手法の研究は、手法自体の研究よりも、実データにより評価し、何が本当に求められているかを検証してみるべき時期にあるように著者は感じている。何より、名義尺度が使えること、得られた知識が人間に可読であること、そして、非線形モデルであることは、従来の重回帰分析には見られない大きな特長である。これら特長を生かして、決定木の適用範囲が、化学・社会学等、多岐に広がってゆくことを念ずるものである。

参考文献

[1] 秋葉他,「例からの学習技術の応用に向けて（1及び2）情報処理,Vol.39,No.2-3.

[2] L.Breiman, et. al,"Classificaion and Regression Trees", Wadsworth,Inc.,1984.

[3] Quinlan 著、古川訳,「AIによるデータ解析」トッパン、1995.

[4] 大滝他,「応用2進木解析法」日科技連、1998.

[5] http://www.salford-systems.com/

[6] http://www.rulequest.com/

[7] B.Efron et. al,"A Introduction to the Bootstrap",Chapmann & Hall,1993.

[8] 金田重郎 http://www1.doshisha.ac.jp/~skaneda

[9]尚、S-PLUSだけではなく、主要な統計解析ツールには、近年、決定木ツールが追加されている。これらの詳細は販売元に問い合わせされたい。