連載講座

ORと統計の融合

数理計画法を用いた最適線形判別関数(5)
—決定木分析との比較—

新村　秀一

1. はじめに

最近，データマイニングがビジネス分野で流行している．色々な手法がある中で，決定木分析（Decision Tree Analysis）は，理論的な分かりやすさと，従来の統計手法の判別分析やクラスター分析と同じ判別や分類を行なう手法であり，応用面が広い手法である．

一方，かつてのAIブームのように，ことさら新しい手法であるという間違った宣伝と，それによる誤解があるようだ．一番大きな問題点は，従来の数理統計学の分かりにくさを逆手にとって，ことさらデータマイニングが従来の統計手法とは異なったアプローチであるかのような説明である．

一方，筆者は統計手法の分類において，基本統計量や相関分析や回帰分析を行った後で，決定木分析を利用すべき手法であるかのように位置づけてきた．しかし，対象データに関して事前の知識や情報が少ない場合，決定木分析を用いて最初に探りを入れ“知の発見を行ない”，その後に本格的なデータ解析を行なうのも良いかもしれないと考えている．

今回は，アヤメのデータとCPDデータを，決定木分析を用いて分析する．そして，その結果をすでに得られているIP-OLDFの判別結果と比較することにする．

IP-OLDFのような新しい手法は，その手法の開発者が，色々な性格の異なったデータを用いて既存の手法と比較評価を行なうべきである．その意味で，今回はIP-OLDFを決定木分析と比較することに主眼がある．ただし，決定木分析には複数の手法があるが，それらを比較したり他の手法との比較研究が少ないので，今回は決定木分析の諸手法の比較評価にもなっている．

しんむら　しゅういち
成蹊大学 経済学部
〒180-8633 武蔵野市吉祥寺北町 3-3-1

2. 決定木分析の概略

決定木分析は，説明変数の値で逐次ケースを枝分かれ状に細分化していき（これが決定木と呼ばれる理由），最終的に幾つかのグループに分ける手法である．最終的なグループは，目的変数が量的変数の場合，目的変数の値の大きなものを含むグループから小さなグループにうまく分類（判別）される．これを「回帰木」という．手法としては，分散分析を繰り返し適用していくイメージになる．目的変数が質的変数の場合，分類されたグループがどちらのカテゴリーになるかがより明らかになる．これを「分類木」という．手法としては，2重クロス集計を繰り返し適用していくイメージになる．

2.1 なぜ重要か

決定木分析は，次のような特徴がある．

- 決定木分析の応用分野は広く，役に立つことである．クレジットカードやスーパーでの顧客を，購買金額の高額なグループから小額なグループに分類できれば，マーケッティングに役立つ．実際にクレジット会社や通販会社においても，高額購買者向けに，お金のかかったカタログを選別的に送ることで費用対効果を上げている．
- 決定木分析は，アルゴリズムが分かりやすく，結果が理解しやすい．一方，これまでの統計手法以上に膨大な計算を行なうコンピュータに適した手法である．
- 決定木分析に限らず，データマイニングのソフトは，一般的な統計ソフトに比べて高額である．しかし，決定木分析は高価なソフトを買わなくても，回帰木は分散分析で，分類木はクロス集計で手間はかかるが分析できる．すなわち，決定木分析は従来のデータ解析の延長線上にある．本研究は，その意味で判別手法と決定木分析の相互比較を目的にしている．
- 決定木分析は，判別分析やクラスター分析という従

来の統計手法と似たような分類を別のアプローチで行なう手法である．

・得られた決定木すなわちDecision Treeは，IF文でもって表現できるので，分析結果を現実の問題に適用するためのシステム化がしやすい．また，医療における診断論理や，AI（Artificial Intelligence，人工知能）とも関係している．

一方，決定木分析の問題点は以下の通りである．

・複数の代表的な手法があるが，それらの比較評価が不充分である．
・分岐を停止する停止則に関する研究が不充分である．
・手法や停止則の組み合わせで，結果が他の手法以上に異なってくる．

2.2 決定木分析のアルゴリズム—AIDとCHAID—

決定木分析は，古くからAID（Automatic Interactive Detector）としてマーケッティング分野で知られた手法を発展させたものである．消費者を購買金額の多い層と少ない層に分けて，木目細かい対応を行なおうというわけだ．

筆者は，『統計・OR活用事典（東京書籍）』でAIDを紹介した際，面白いが制約が多いと指摘している．AIDは，目的変数が量的変数で，説明変数が性別や喫煙の有無のように2値の値に限定されているからである．すなわち，ケースは逐次的に2分岐される．なぜ，計算時間がかかりアルゴリズムが多少複雑になるにしたとしても，3カテゴリー以上の多分岐にしないのだろうかと不思議に思った．問題意識があれば，自分で改良すればよいのに，私にはその才能がないので，それ以上のことをしなかった．

その後数年して，目的変数も説明変数も量的と質的の両方が扱え，多分岐するCHAIDのことを知った．Chisquared-AIDの略である．AIDは，分散分析を逐次適用していくのに対して，CHAIDは目的変数と説明変数の2重クロス集計を逐次行なっていくイメージである．クロス集計を用いることで，多分岐が可能になった．

ただし，目的変数が量的変数であれば，AIDと同じく分散分析が用いられる．

CHAIDの他にも幾つかの手法が開発され，最近ではそれらをまとめてデータマイニングという新しい革袋の中で，決定木分析という中核的な手法として位置付けられている．

3. CPDとアヤメのデータを決定木分析する

3.1 AnswerTreeの紹介

ここでは，決定木分析にSPSSのAnswerTreeを用いる．AnswerTreeには，CHAID（Chi-squared Automatic Interactive Detector）の他に，より探索を綿密に行なうExhaustive-CHAIDと，2分岐に限定したC＆RT（CART）とQUESTの4つの手法がある．C＆RTは，Classification and regression treesの略で，分類木と回帰木を行なう手法である．QUESTは，Quick, Unbiased and Efficient Statistical Treeの略で，分類木を高速で行なう手法である．

3.2 CPDの停止則による結果の違い

(1) 停止則

決定木分析には，停止則として，分岐する階層の最大数，親ノードと子ノードに含まれるケースの最小数，そしてχ^2検定やF検定の有意水準がある．

分岐する階層の最大数は，小さくすると無条件で停止する強い制約力をもっている．そこで300件程度のデータに充分と思われる5階層にする．また有意水準は，ケース数が少ない今回の場合，デフォルトの5%のままにする．

表1は，停止則による結果の違いを示す．最初の4個（No. 1～No. 4）はCHAID，次の4個（No. 5～No. 8）はExhausive-CHAID，次の4個はC＆RT，最後の4個はQUESTである．階層は全て5に固定し，親ノードと子のノードに含まれるケースの最小数を同じ値にして，上から順に20個，10個，5個，1

表1 CPDの停止則の変更による結果

No	Stopping Rule Node	Results Level	T-Node	Error	Var.	OLDF Error	Var.
1	20	2	6	35	7,10,12,15	10	9,12,15,18
2	10	2	8	18	7,12,16	12	9,12,18
3	5	3	8	15	7,12,18	12	9,12,18
4	1	5	16	7	2,5,7,9,12 15,16,18,19	4	1,2,5,8,9 12,15,17,18
5	20	2	5	30	4,12	13	9,12
6	10	2	7	25	4,12	13	9,12
7	5	2	7	25	4,12	13	9,12
8	1	3	10	20	1,4,9,12,16	8	2,9,12,15,18
9	20	2	3	20	12	19	12
10	10	2	4	17	9,12	13	9,12
11	5	4	7	9	9,12,15,18	10	9,12,15,18
12	1	5	16	1	2,5,6,9,10, 12,15,18,19	4	1,2,5,8,9 12,15,17,18
13	20	2	3	22	12	19	12
14	10	3	4	22	12	19	9,12
15	5	4	6	18	5,12,18	12	9,12,18
16	1	5	14	8	1,2,5,9 12,15,17,18	6	1,2,8,9,12 15,17,18

個と減少させた．本来は，子ノードは親ノードよりも少なくすべきであるが，その組み合わせによって結果が大きく左右され，基準がないので同じ値を用いる．有意水準は全てソフトウエアのデフォルトに固定してある．

1番目のCHAIDの結果は，階層が5で，親ノードと子ノードが20と20，χ^2検定が5%である．その結果，2階層（Level）でターミナルノード（T-Node）が6個，採用されている説明変数（Results欄のVar.）がX7，X10，X12，X15の4個で，誤分類数が35個（誤分類率35/240=0.126）である．

一方，IP-OLDFのモデル選択で選ばれた4変数（X9，X12，X15，X18）の誤分類数は10個（誤分類率10/240=0.042）である．

誤分類数は，決定木分析の方が25個（10.4%）多いことになる．

(2) 比較結果

同じ条件の決定木分析とIP-OLDFを比較すると，11番目と12番目のC&RTだけが，わずかに決定木分析の方がIP-OLDFより誤分類数が少ない．

一方，各手法ごとにノード数を減らしていくと，決定木分析の誤分類数は減少する．それに対応して変数の個数は，1番目のCHAIDを例外とすれば，選ばれる説明変数は単調に増加していく傾向がある．

以上から，決定木分析の親子のノード数に関して，少なくしていけば一般的に選ばれるターミナルノード数と説明変数が増え，その結果見かけの誤分類数が少なくなる傾向があるようだ．

決定木分析は，IP-OLDFと比較して誤分類数が多く，客観的な停止則の選択が難しいといえる．

決定木分析の手法に関しては，説明変数の個数は別途考慮する必要があるが，C&RTの誤分類数が一番少なく，次にQUESTになる．No.1とNo.5を別とすれば，3番目がCHAIDであり，さも探索が綿密に行なわれるExhausive-CHAIDの成績が一番悪いことは，注目に値する．

この点に関しては，ソフトウエアが高価なため実際のデータで確かめないで，これまでカタログ機能だけでExhausive-CHAIDが良いと判断してきた不明を恥じるばかりである．

しかし，このような比較評価は手法の開発者が事前に十二分に行なうべきものであろう．

3.3 アヤメの停止則による結果の違い

表2は，アヤメのデータにおける停止則による結果の違いを示す．最初の4個はCHAID，その後はExhausive-CHAID，C&RT，QUESTである．階層は全て5に固定し，親ノードと子のノードを同じ値にして，上から順に20個，10個，5個，1個と減少させた．有意水準は全てデフォルトに固定してある．

表2 アヤメの停止則による結果

No	Stopping Rule Node	Results Level	T-Node	Error	Var.	OLDF Error
1	20	1	3	7	4	5
2	10	1	3	7	4	5
3	5	2	4	7	3,4	3
4	1	2	4	6	3,4	3
5	20	1	3	7	4	5
6	10	1	7	7	4	5
7	5	1	7	7	4	5
8	1	2	9	5	2,3,4	2
9	20	2	3	6	3,4	3
10	10	2	3	6	3,4	3
11	5	3	4	4	3,4	3
12	1	5	9	0	1,3,4	3
13	20	2	3	6	3,4	3
14	10	3	4	6	3,4	3
15	5	3	4	6	2,3,4	2
16	1	5	9	3	2,3,4	2

最初のCHAIDは，階層が5で，親ノードと子ノードが20と20，χ^2検定が5%である．その結果，1階層でターミナルノードが3個，採用されている説明変数がX4（花びら幅）の1個で，誤分類数が7個である．ここでは，がく片をX1，がく片幅をX2，花びらをX3，花びら幅をX4とする．

一方，IP-OLDFのX4によるモデルの誤分類数は5個である．

誤分類数は，決定木の方が2個多いことになる．

12番目のQ&RTだけが，決定木分析の方がIP-OLDFより誤分類数が3個少ない．しかし，親子のノード数が1で，5階層の3変数モデルであり，もともと1変数か2変数で十分判別できることが分かっているので，現実において採用が難しい．

各手法ごとにノード数を減らしていくと，決定木分析の誤分類数は減少する．選ばれる説明変数の個数は，増加していく傾向がある．

以上から，決定木分析のノード数に関して，少なくしていけば選ばれるターミナルノード数と説明変数が増え，その結果見かけの誤分類数が少なくなる傾向があるようだ．

4手法を比較すると，Q&RTとQUESTが，Exhausive-CHAIDとCHAIDよりわずか1〜2例の違いであるが成績が良く，Q&RTやQUESTの方が説明変数がたかだか1個多い場合がある．

4. CPDのCHAIDによる分類木

CPDデータに関して，表1では自然分娩群180例をA群，帝王切開群60例をB群とし，4手法を用いて分類木により分析した．

4.1 分析結果

図1は，CPDデータをExhaustive-CHAIDで分析した表1のNo.5の実行結果である．

目的変数を自然分娩群と帝王切開群の2群として，説明変数としてX1からX19の変数を指定した．

分岐する階層の最大数を5として，親ノードの個数を20そして子ノードの個数を20に指定して，それ以下になると停止する規則を用いて分析した．

240人の患者（GROUP）は，最初の分岐でX4を92＜X4＜＝103，103＜X4＜＝113と113＜X4＜＝150の3群に分割するのが，他の変数で分割するより，よく違いを表わせることを示す．

GROUPはルートノードと呼ばれ，X4で分岐した3つのノードはその子ノードになる．親ノードの240人が，子ノードの27人と86人と127人に分割された．いずれのノードも20人以上である．

次に，X4で分けられた最初のノードは，親ノードが40以下なので子ノードが20以下になり分岐できず停止し，ターミナルノードになる．残りの2つの子ノードを親ノードとして，分析が行なわれる．103＜X4＜＝113のノードは，86人がX12でもって300以下とそれ以上の2つのノードに分かれ，それらがターミナルノードになる．

X4が113＜X4＜＝150のノードは，これを親ノードとして，X12でもって300以下とそれ以上の2つのノードに分かれ，それらがターミナルノードになる．

結局，240人の患者は5つのターミナルノードに分割された．

CHAIDやExhausive-CHAIDの便利な点は，最適な分割を自動的に行なってくれる点である．図1のX4とX12の分岐はソフトウエアが探してくれる．

ただし，No.6で親子のノード数を20から10にすると，ノード2の第2層のX12は127と300で3分岐し，ノード3の第2層のX12は228と300で3分岐して，7群に分かれる．

また同じ条件で，CHAIDとExhausive-CHAIDを比較すると，CHAIDの方が分岐の数が多くなり，誤分類数が少なくなるのは納得できない．Exhausive-CHAIDは，CHAIDに比べてより最適な分岐を詳しく探すことが特徴であるので，より少ないターミナルノードで誤分類数も少ないルールが得られるものと期待するのが当然である．しかし，そのような結果にならないようだ．

4.2 ルール

図2は，ターミナルノードを選別するルールである．一般のIF文の形式であるが，SPSSやSQL形式のルールも出力できる．

最初のIF文で，X4＜＝103の条件を満たす27人が選ばれる．27人中23人が帝王切開群（B群）なので，予測値はbと表示され，bが選ばれる確率が0.852（＝23/27）であることが分かる．そして，これをノード1としている．

ノード4～7のルールも同様である．

これらのルールを利用して，External Checkや現実のシステムへの適用ができる．

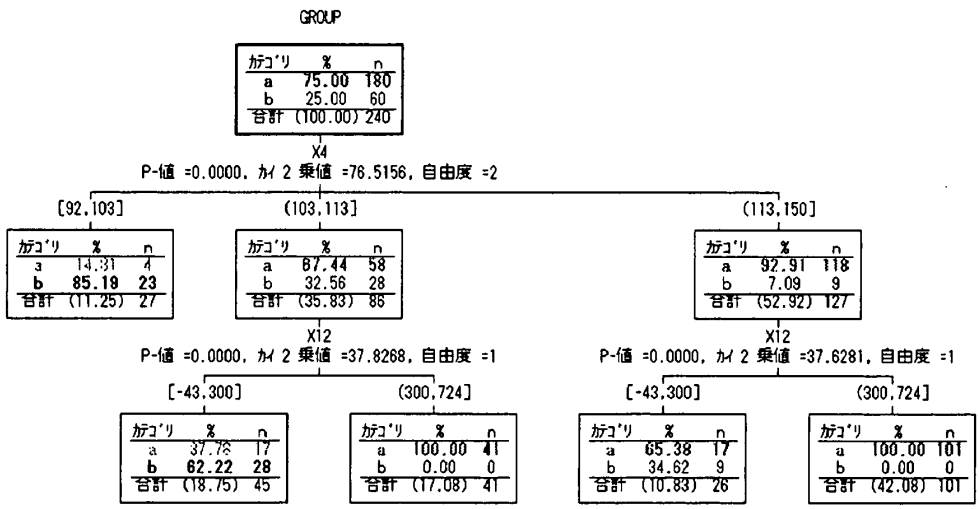

図1 CPDのNo.5の分類木の結果

```
/* ノード 1*/
IF (x4 NOT MISSING  AND  (x4 <= 103))
THEN ノード = 1  予測値 = 'b '  確率 = 0.852
/* ノード 4*/
IF (x4 NOT MISSING  AND  (x4 > 103  AND  x4 <= 113))
AND  (x12 IS MISSING  OR (x12 <= 300))
THEN ノード = 4  予測値 = 'b '  確率 = 0.622
/* ノード 5*/
IF (x4 NOT MISSING  AND  (x4 > 103  AND  x4 <= 113))
AND  (x12 NOT MISSING  AND  (x12 > 300))
THEN ノード = 5  予測値 = 'a '  確率 = 1.000
/* ノード 6*/
IF (x4 IS MISSING  OR (x4 > 113))
  AND  (x12 NOT MISSING  AND  (x12 <= 300))
THEN ノード = 6  予測値 = 'a '  確率 = 0.654
/* ノード 7*/
IF (x4 IS MISSING  OR (x4 > 113))  AND  (x12 IS MISSING  OR (x12 > 300))
THEN ノード = 7  予測値 = 'a '  確率 = 1.000
```

図2　Exhausive-CHAIDのルール

表3　CPDのNo.5の応答表

応答の要約
目的変数: GROUP　　カテゴリ目的変数:

ノードごと ノード	ノード: n	ノード: %	正答数: n	正答率: %	応答率 (%)	インデックス (%)
1	27	11.25	23	38.33	85.19	340.74
4	45	18.75	28	46.67	62.22	248.89
6	26	10.83	9	15.00	34.62	138.46
7	101	42.08	0	0.00	0.00	0.00
5	41	17.08	0	0.00	0.00	0.00

4.3　ターミナルノードの評価

表3はターミナルノードを序列化する応答表である．序列化は，ターミナルノードに含まれる帝王切開（B群）に注目して行なわれる．

ノード1のケース数は27人で，240人に対して11.25%である．ノード1には23人のB群が含まれその例数が正答数の列に表示される．B群全体の60人に対して38.33%である．応答率の85.19%は，23/27のことである．インデックスの340.74%は，応答率をB群の全体での比率60/240で割った値である．すなわちノード1には，全体のB群の比率の3.4倍にあたる帝王切開群の患者が含まれていることを表わす．

ターミナルノードの序列化は，このインデックスの大小順になる．ノード7と5のインデックスは0であるが，その場合例数の多いほうの序列が上になる．

4.4　分類結果の評価

表4は，図1の分類結果である．事前確率を考えないで，例数の多いほうに判別した結果が判定の列に示されている．

これを誤分類行列にまとめたものが表5である．A群の180例は，159例が正しくA群と分類され，21例がB群と誤分類された．B群の60例は，51例が正しくB群と分類され，9例がA群と誤分類された．12.5%（=30/240）が誤分類確率になる．

表4　CPDのNo.5の分類結果

ノード	A群	B群	判定
1	4	23	b
4	17	28	b
6	17	9	a
7	101	0	a
5	41	0	a

表5　CPDのNo.5の誤分類行列

		実際のカテゴリ a	b	合計
予測されたカテゴリ	a	159	9	168.00
	b	21	51	72.00
	合計	180	60	240.00

5.　分類木による問題点

5.1　CPDのC&RTによる良いモデル

図3は，表1の10番目のC&RTによる分析結果である．図1に比べ同じ2階層であるが，C&RTは2分岐に制限されているので第1階層も第2階層も2分岐し，4つのターミナルノードが選ばれている．これによって，IP-OLDFで2変数の最良モデルとしたX9とX12の同じ組が選ばれた．

表6は，分析結果の応答表である．表7は誤分類表である．17例が誤分類され，見かけの誤分類確率は7%になる．モデルとしてもCHAIDに比べて良さそうだ．

5.2　停止則の問題点

図4は，CPDデータで一番誤分類数が少なかったC&RT（表1の12番目）の分類木の樹木図である．

表8は，それに対応する応答表である．ノード5からノード6にB群のデータが含まれ，ノード30以下にはA群のデータしかないことが分かる．しかも，ノード5（第5層）に60人中42人のB群が，ノード30（第2層）にA群の160人中154人のデータが含まれ，その他のノードのデータ数が少ないことが分かる．単に9次元のデータ空間で，内部標本にあわせてデータを誤分類数を最小にするように判別したモデルなので，その結果を外部標本に適用することが困難であろうことは容易に想像できる．

以上見たように，親子のノード数を減少すると，一般的にターミナルノードや階層や説明変数が多くなり，

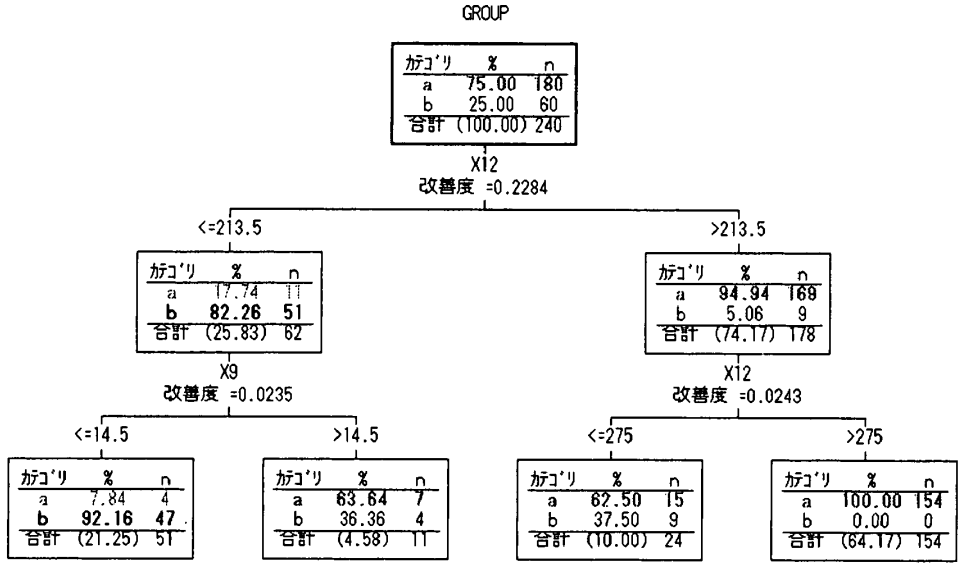

図3　C＆RT（表1のNo. 10）の決定木

表6　C＆RT（表1のNo. 10）の応答表

目的変数: GROUP　　ｶﾃｺﾞﾘ目的変数: b

ﾉｰﾄﾞごと

ﾉｰﾄﾞ	ﾉｰﾄﾞ: n	ﾉｰﾄﾞ: %	正答数: n	正答率: %	応答率 (%)	インデックス (%)
3	51	21.25	47	78.33	92.16	368.63
5	24	10.00	9	15.00	37.50	150.00
4	11	4.58	4	6.67	36.36	145.45
6	154	64.17	0	0.00	0.00	0.00

表7　C＆RT（表1のNo. 10）の誤分類表

		実際のｶﾃｺﾞﾘ a	b	合計
予測されたｶﾃｺﾞﾘ	a	176	13	189
	b	4	47	51
	合計	180	60	240

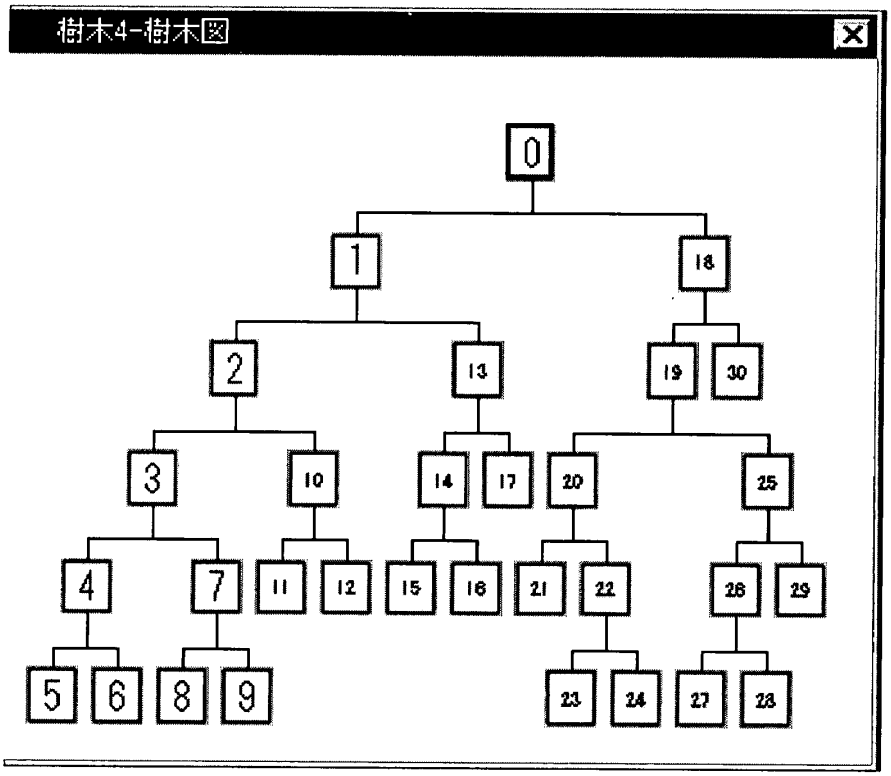

図4　表1のNo. 12の樹木図

表8　表1のNo.12の応答表

ノード	ノード:n	ノード:%	正答数:n	正答率:%	応答率(%)	インデックス(%)
5	42	17.5	42	70.00	100	400
21	5	2.08	5	8.33	100	400
16	4	1.67	4	6.67	100	400
28	3	1.25	3	5.00	100	400
24	1	0.42	1	1.67	100	400
9	1	0.42	1	1.67	100	400
11	1	0.42	1	1.67	100	400
6	4	1.67	3	5.00	75	300
30	154	64.17	0	0.00	0	0
29	11	4.58	0	0.00	0	0
17	6	2.5	0	0.00	0	0
27	2	0.83	0	0.00	0	0
12	2	0.83	0	0.00	0	0
23	2	0.83	0	0.00	0	0
8	1	0.42	0	0.00	0	0
15	1	0.42	0	0.00	0	0

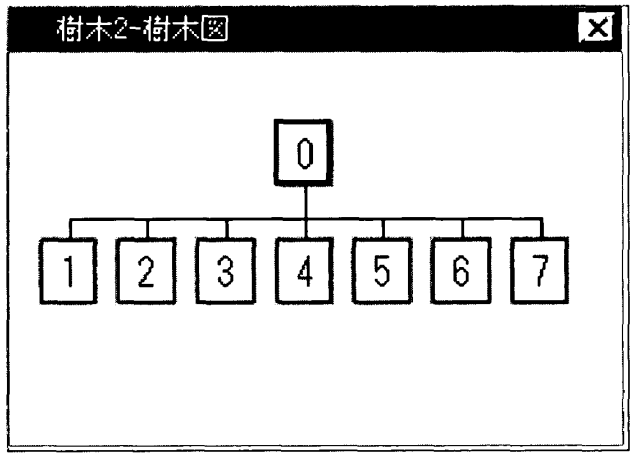

図5　アヤメのCHAIDの樹木図

その結果見かけの誤分類数は少なくなる．これは，回帰分析において決定係数がモデル選択に役立たないのと似たような状況である．AICに似たような別の停止規則が必要であろう．

データマイニングにおいて，集めたデータをモデル作成用，枝刈り用，評価用に3分割することが提案されている．すなわち，モデル作成用のデータで決定木を求め，それを枝刈り用のデータで不用と思われる枝を刈りターミナルノードを減らす．その後，評価用データで誤分類数を計算し評価に用いるという3段階の手順を踏むことである．これを行なうには，相当数のデータがなければ実現困難であり，手順そのものが恣意的であり煩雑という欠点がある．

小標本に対しては，別途停止則を含めて考える必要があろう．回帰木では，ターミナルノードを多重比較で検討し，枝刈りすることを提案している（新村・新村）．

5.3　CHAIDの問題点

筆者は，AnswerTreeのソフトが高額のため，文献レベルで制限の少ないExhaustive-CHAIDが一番分析結果が良いと長らく考えてきた．しかし，今回の比較評価で思いのほか，他の2分岐に限定された手法に比べて誤分類数は多かった．この大きな理由として，上の階層で2分岐に限定されず3分岐以上に分かれることが原因と考えられる．例えば，図5は表2の5番目の樹木図である．第1階層で7分岐し，それ以上の分岐が行なえなくなり停止している．

筆者が社会人になって最初に師事したのが大阪成人病センターの野村裕医師である．同氏は，日本の心電図の自動診断の先覚者であり，枝分かれ診断の権威であった．同氏の研究成果の1つは，単純に枝分かれで診断しては実用化できず，あるものは上の階層にフィードバックする必要があるという点である．上の階層で，局地最適解を得たとしても，それが大域的な最適解になっていないことと似た状況であろう．

6.　結論

今回用いたCPDとアヤメのデータでは，判別結果や解の安定性に関して，IP-OLDFの方がはるかに成績が良いことが分かった．

一方，決定木分析の手法に関してはExhausive-CHAIDは，分岐に関する制約が少なく，計算時間がかかるが探索を綿密に行なう点から，好成績が期待されたが予想外の成績の悪さであった．

AnswerTreeの購入は，成蹊大学研究助成の賜物である．

参考文献

[1] 新村秀樹，新村秀一（2002），決定木分析のモデル選択に関する検証(1)，2002年度春季大会．

[2] 森村英典，牧野都治編（1984），統計・OR活用事典，東京書籍．