2003年日本オペレーションズ・リサーチ学会
春季研究発表会

決定木分析のモデル選択に関する考察（3）

—科学万博データによる考察—

01207730　ムトーテクノサービス　　＊新村秀樹　SHINMURA　Hideki
01202720　成蹊大学　　新村秀一　SHINMURA　Shuichi

1. はじめに

2002年春季・秋季研究発表会において「BANK.SAV」という474件のデータを用いて、「決定木分析のモデル選択に関する考察（1），（2）」という題で発表を行った（文献[1]，[2]）。

文献[1]では、大規模データによる枝刈（交差妥当化）は、小規模データに用いることができないので、それに替わる「多重比較による枝刈」の提案、等を行った。

文献[2]では、決定木分析（Answer Tree）の4手法（CHAID、Exhaustive CHAID、C&RT、QUEST）の比較を行った。

今回は特に、文献[2]で行った決定木分析の4手法の比較を別のデータで行った。

2. BANK.SAVで得られた知見

BANK.SAVで決定木分析の4手法の比較を行い、得られた知見は次の通りである。

・Exhaustive CHAIDはCHAIDと比較して予想外に成績がよくない
・C&RTは誤分類では最もよい成績が得られた
・QUESTは、C&RTを用いることができるのであれば、用いる必要はない
ただし、
・C&RTとQUESTは極端に件数の多いものと極端に件数の少ないものに分化させるので注意が必要である。

3. データと手法

今回、分析に用いたデータは「1985年の国際科学技術博覧会（いわゆるつくば科学万博'85）の生データ」である（文献[3]）。全数調査データで、184件のデータである（文献[3]）。

「第1、第3四分位数で3カテゴリ化した入場者」を目的変数とし、「月」、「天気」、「曜日」、「午前天気」、「午後天気」、「降車数」、「分担率」、「団体バス」、「マイカー」、「前日午前天気」、「前日午後天気」、の11変数を説明変数に用いる。

「Answer Tree」の停止則は文献[2]と同じく、樹木の深さ（階層数）は7、親ノード、子ノードに最低含まれるケース数を20、10、5、1、とする。

4. 分析結果

表1は4手法の条件を変えた分析結果である。

最も成績がよいのは、「CHAIDのノード数が20個と5個」、「C&RTのノード数が1個」、「QUESTのノード数が10個」、である。最も成績がよくないのは、「Exhaustive CHAIDのノード数が10個、5個、1個」、「C&RTのノード数が10個」、「QUESTのノード数が20個」、である。

階層数、ターミナルノード数、分析に用いた変数の数は、停止則が緩くなるほど増加している。誤分類数は停止則が緩くなるほど減少している。第1層での分化は、CHAIDとExhaustive CHAIDでは停止則が緩くなるほど増加するのに対し、C&RTとQUESTでは2で固定されている。これはC&RTとQUESTが「2分岐」であるためである。また、表1において、C&RTのノード数が20個のものと10個のものは得られた結果が全く同じであるが、樹木図も全く同じであった。

文献［2］では、「多分岐」と「2分岐」に分けて考え、2分岐の方が誤分類数は少ないが、極端にデータ件数が多いターミナルノードと1桁のターミナルノードに分化させると指摘した。

今回においてはそのような傾向は見られず、多分岐よりも2分岐の方が、データ件数が1桁のターミナルノードが多い傾向にある、という程度にとどまった。

以上のことから、秋季で得られた知見の中で今回確認できた事柄は、「Exhaustive CHAID は CHAID と比較して予想外に成績がよくない」、ということのみである。

表1　分析結果

	Stopping Rule Node	Level	1st Classification	T-Node	Error	Var.
CHAID	20	3	3	5	39	マイカー、団体バス
	10	3	4	9	33	マイカー、団体バス、Weather AM、午前天気
	5	4	4	12	19	マイカー、団体バス、降車数、Weather AM、午前天気、月
	1	5	4	14	18	マイカー、団体バス、降車数、Weather AM、月、曜日
E-CHAID	20	1	4	4	52	マイカー
	10	2	9	10	41	マイカー、月
	5	2	9	12	34	マイカー、降車数、月、午後天気
	1	3	9	20	21	マイカー、降車数、月、天気、Weather PM、分担率、曜日
C&RT	20	2	2	3	41	マイカー、降車数
	10	2	2	3	41	マイカー、降車数
	5	6	2	10	23	マイカー、降車数、天気、月、団体バス、曜日
	1	7	2	22	8	マイカー、降車数、天気、分担率、曜日、月、団体バス、Weather PM
QUEST	20	1	2	2	61	マイカー
	10	6	2	11	28	マイカー、降車数、天気、月、曜日、団体バス
	5	6	2	16	26	マイカー、降車数、天気、月、曜日、団体バス、Weather AM
	1	7	2	30	13	マイカー、降車数、天気、月、曜日、団体バス、Weather AM

5. 回帰分析による比較

今回はより客観的に検討するため、回帰分析による比較を行う。表2はJMP5.0.1で誤分類数を応答とし、階層の数、第1層の分岐数、ターミナルノード数（TN）、分析に用いられた説明変数の数、CHAID と Exhaustive CHAID と C&RT をダミー変数として、変数増加法によって得られたモデルである。

表2　ステップワイズ法による分析結果

	万博		BANK.SAV(初任給)		BANK.SAV(現在の給与)	
パラメータ	推定値	p値	推定値	p値	推定値	p値
切片	64.1467686	1	116.285524	1	107.608686	1
階層	-1.6473376	0.08492498				
第1層						
TN	-0.6579255	0.00519295	-1.2522202	0.00002328	-1.6274337	0.00092276
変数	-2.7058436	0.00337057			2.78633338	0.1769137
CH	-11.963701	0.0001044	-11.939165	0.02909651	(0	0.69987907)
ECH	-6.8154972	0.0195315	-10.00222	0.05956385	8.33425449	0.01219937
CRT	-10.468995	0.00023909	-23.434725	0.00045754	-12.972554	0.00052448

「万博」では、「第1層」以外は全てモデルに用いられた。CHAID が C&RT よりもわずかによい成績である。Exhaustive CHAID は CHAID よりよくない成績である。

「BANK.SAV（初任給）」では、ターミナルノードと3個のダミー変数が用いられた。C&RT が抜きんでてよい成績である。Exhaustive CHAID は CHAID よりよくない成績である。

「BANK.SAV(現在の給与)」では、ターミナルノード、変数、C&RT、Exhaustive CHAID、の4つが用いられた。変数が正の値となっており、決定木の結果とは反している。これは変数とターミナルノードとの相関が強いためと推測される。つまり、ターミナルノードが効きすぎているために補正をしているのであろう。他の2つとは異なり、CHAID が回帰係数として選ばれなかった。

6. まとめ

「Exhaustive CHAID は CHAID と比較して予想外に成績がよくない」ことは回帰分析でも確認することができた。また、誤分類数はターミナルノード数が多い程少なくなることが明らかになった。

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考察（1），2002 年春季研究発表会アブストラクト集，pp.142-143.

[2] 新村秀樹，新村秀一(2002)，決定木分析のモデル選択に関する考察（2），2002 年秋季研究発表会アブストラクト集，pp.232-233.

[3] 新村秀一(1989)，『易しく実践　データ解析の進め方』，共立出版