2002年日本オペレーションズ・リサーチ学会
秋季研究発表会

決定木分析のモデル選択に関する考察（2）

—4手法の比較評価—

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

2002年春季研究発表会において、決定木分析のモデル選択に関し、多重比較による枝刈による方法を検討した。今回は決定木分析の4手法の手法について比較・評価を行いたい。

1. データと手法

データは前回と同じく、「米国銀行員のデータ」である。手法はAnswer Treeに含まれる4手法（CHAID、Exhaustive-CHAID、C&RT、QUEST）を用いる。BANK.SAVには「初任給」と「現在の給与」という変数がある。これらを目的変数として誤分類率などで4手法を比較・評価したい。

QUESTは、目的変数が名義変数に制限された手法である。この制限に合わせるため、「初任給」と「現在の給与」を第3四分位数と第1四分位数を参考にして3分割し、値の高いカテゴリから低いカテゴリまで順に3、2、1とした新変数を作成し、4手法を同じ条件で分析する。

「初任給」を分析する場合の説明変数は、「現在の給与」を除く、「職種」、「性」、「人種」、「性・人種」、「熟練度」、「年齢」、「就学年数」、「就業年数」、の8変数を用いる。

「現在の給与」を分析する場合には、「（新変数ではない）初任給」を加えた9変数を用いる。

決定木分析では停止則という、決定木の樹木の成長をどの段階で止めるかを決めるオプションがある。今回は各手法の比較・検討であるので、これを揃えなければならない。樹木の深さ（階層数）はいずれの場合も「7」を上限とする。これは、階層数が少ないと十分な樹木の成長のないまま打ち切られるからである。親ノードと子ノードの最小ノード数は固定し、「20」、「10」、「5」、「1」、とする。本来、子ノードは親ノードよりも少なくすべきであるが、これを定める基準がないので固定することにした。つまり各手法で8通り（ノードの4つの停止則＊2つの説明変数）の出力が得られる。

2. 分析結果

表1は初任給の結果である。

最も成績がよいのはC&RTであり、最も成績が悪いのはQUESTとノード数が1個のExhaustive CHAIDである。C&RTとQUESTは共に2分岐する手法であるが、結果は正反対となっている。CHAIDとExhaustive CHAIDでは、前者よりも後者の方が結果はよいと思われたが、必ずしもそうとはいえないようである。ノード数が5個と1個ではCHAIDの方がよくなっている。

表1　手法ごとの結果（初任給）

	Stopping Rule Node	Level	T-Node	Error	Var.
CHAID	20	3	10	101	(性別, 職種, 就学年数, 人種, 就業年数)
	10	3	11	91	(性別, 職種, 就学年数, 人種, 就業年数)
	5	4	16	76	(性別, 職種, 就学年数, 人種, 就業年数, 年齢, 仕事の習熟度)
	1	6	24	73	(性別, 職種, 就学年数, 年齢, 就業年数, 人種, 仕事の習熟度)
Exhausive-CHAID	20	3	10	97	(性別, 就学年数, 就業年数)
	10	3	12	91	(性別, 職種, 就学年数, 性別と人種, 就業年数)
	5	4	14	83	(性別, 職種, 就学年数, 性別と人種, 就業年数, 仕事の習熟度)
	1	5	24	79	(性別, 職種, 就学年数, 年齢, 就業年数, 人種, 仕事の習熟度, 性別と人種)
C&RT	20	5	6	94	(職種, 性別, 年齢)
	10	6	9	85	(職種, 性別, 就学年数, 年齢, 性別と人種, 就業年数)
	5	7	18	63	(職種, 性別, 就学年数, 年齢, 仕事の習熟度, 人種, 就業年数, 性別と人種)
	1	7	36	43	(職種, 性別, 就学年数, 年齢, 仕事の習熟度, 人種, 就業年数, 性別と人種)
QUEST	20	3	5	104	(就学年数, 性別, 性別と人種)
	10	3	6	104	(就学年数, 性別, 性別と人種, 人種)
	5	7	17	99	(就学年数, 性別, 性別と人種, 職種, 人種, 仕事の習熟度, 年齢)
	1	7	36	79	(就学年数, 性別, 性別と人種, 職種, 人種, 就業年数, 仕事の習熟度, 年齢)

誤分類数は、停止則がゆるくなるほど減少している。また、停止則がゆるくなるほど、説明変数の数は増加している。

表2は現在の給与の結果である。

表2　手法ごとの結果（現在の給与）

	Stopping Rule Node	Level	T-Node	Error	Var.
CHAID	20	5	11	111	(性別, 初任給, 就学年数, 年齢, 人種)
	10	4	16	98	(性別, 初任給, 職種, 就業年数, 就学年数, 仕事の習熟度)
	5	5	19	86	(性別, 初任給, 職種, 就業年数, 就学年数, 仕事の習熟度)
	1	6	26	80	(性別, 初任給, 職種, 就業年数, 就学年数, 仕事の習熟度)
Exhausive-CHAID	20	4	13	111	(性別, 初任給, 就学年数, 年齢, 人種)
	10	4	16	106	(性別, 初任給, 職種, 年齢, 就学年数, 性別と人種)
	5	4	18	101	(性別, 職種, 就業年数, 就学年数, 初任給, 年齢)
	1	6	27	95	(性別, 職種, 初任給, 就業年数, 就学年数, 年齢, 仕事の習熟度, 人種)
C&RT	20	4	7	93	(職種, 初任給, 年齢)
	10	5	11	88	(職種, 初任給, 年齢, 就業年数)
	5	7	17	81	(職種, 初任給, 年齢, 就業年数, 仕事の習熟度)
	1	7	40	53	(職種, 初任給, 年齢, 仕事の習熟度, 性別と人種, 性別, 就業年数, 人種, 就学年数)
QUEST	20	4	5	107	(初任給, 年齢)
	10	6	10	99	(初任給, 職種, 年齢, 性別)
	5	6	16	96	(初任給, 職種, 年齢, 就業年数, 性別, 性別と人種)
	1	7	38	77	(初任給, 職種, 年齢, 性別, 仕事の習熟度, 職種, 就学年数, 就業年数, 性別と人種)

最も成績がよいのは初任給と同じくC&RTであり、よくないのはExhaustive CHAID全てとCHAIDのノード数が20個の場合である。現在の給与でもCHAIDよりもExhaustive CHAIDの方が成績はよくない。初任給の場合よりも差は広がっている。

誤分類数は、初任給がゆるくなるほど減少している。ターミナルノード数でも説明変数でも初任給と同じく、停止則がゆるくなるほど増加しているが、階層数ではCHAIDの停止則のノード数が10個で減少している。階層数はこの場合以外は、やはり増加傾向にあるといえる。

3. まとめ

結果を見る限り、「多分岐（CHAIDとExhaustive CHAID）」と「2分岐（C&RTとQUEST）」に分けて考えた方がよさそうである。

初任給ではそれ程ではないが、現在の給与では、C&RTのノード数が20個の誤分類数は93であり、他の手法は111～107であるので違いは明らかである。同じく2分岐のQUESTは、初任給ではどの場合でも最下位であるが、現在の給与のノード数20個では、CHAIDとExhaustive CHAIDよりも誤分類数は少ない。この2分岐の誤分類数の少なさは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いう性質がある。しかし、決定木分析の目的は「統計的に有意義な見解を発見すること」である。一方で極端に多い件数を分類し、他方で極端に少ない件数を分類する手法はこの点に関してはどうなのであろうか。

次に「同じ変数が繰り返し出てくる」ことについて述べる。「同じ変数が繰り返し出てくる」ことは「2分岐」のみで確認できた。これが「2分岐」の特徴なのかどうか、また、同じ変数を繰り返し用いることが有用であるかどうかは現時点ではわからない。

今回の4手法の比較から、次のことがいえる。

・C&RTは誤分類では最もよい成績である（誤分類数が最も少ない）
・C&RTとQUESTは類似点が多いが、前者の方が後者よりも優れた成績を収めた。よってC&RTを用いることができるのであれば、QUESTを用いる必要はない

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.