2002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（1）

—多重比較による枝刈—

法政大学社会学部　＊新村秀樹　SHINMURA　Hideki
01202720　成蹊大学　新村秀一　SHINMURA　Shuichi

最近、データマイニングの「決定木分析」が知識発見の手法として注目されている(文献1)。そこで、学生の立場で知識発見が容易かどうか、そして停止則(枝刈)に関して検討を行なった。

1　データと手法

BANK.SAV は、「雇用機会均等訴訟に関係する米国の銀行の 1969 年から 1971 年までの 3 年間に雇用された 474 人の行員データ」である（文献 2)。今回このデータをエス・ピー・エス・エス(株)の Website（http://www.spss.co.jp/）より入手し分析する。

SPSS の Answer Tree には、CHAID、Exhaustive-CHAID、C&RT、QUEST の 4 手法がある。機能が良いと思われる Exhaustive-CHAID を用いて決定木分析を行なう。

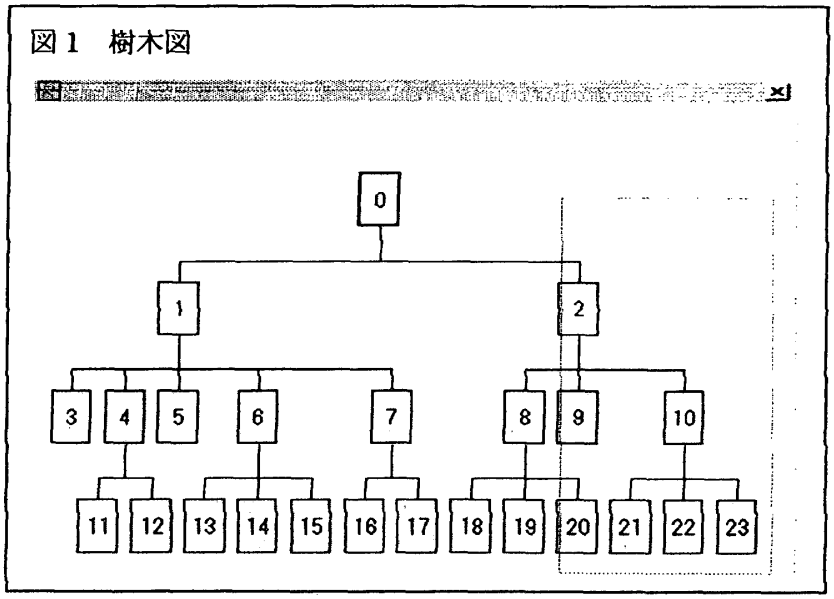
「初任給」を従属（目的）変数とし、「従業員コード」と「現在の給与」を除く「職種」、「性」、「人種」、「性・人種」、「熟練度」、「年齢」、「就学年数」、「就業年数」、の 8 変数を独立（説明）変数とする。停止則として階層は 3、親ノードのケースは 10、子ノードのケースは 5、とする。これは何度かの試行の末に決めた。

2　分析結果

2－1　樹木図と層別箱ヒゲ図

図 1 は初任給を目的変数とした樹木図である。3、5、9、11 から 23 の 3 階層 16 個のターミナルノードに分かれた。

図 1　樹木図



この 16 個のターミナルノードを作成するルールを作成し、図 2 のように 16 個のターミナルノードによる層別箱ヒゲ図を作成した。ターミナルノード 3 は、初任給の平均値の一番高いグループである。上にある丸印は、大きな外れ値をあらわす。ターミナルノード 11 と 12、13 から 15、16 と 17、18 から 20、21 から 23、という組み合わせは同じ親を持つ子ノードである。

ターミナルノード 11 と 12、13 と 14、18 と 19、21 から 23、の組み合わせは重なっているので枝刈して、一つのノードにまとめても良さそうである.

2—2　停止則の改善点

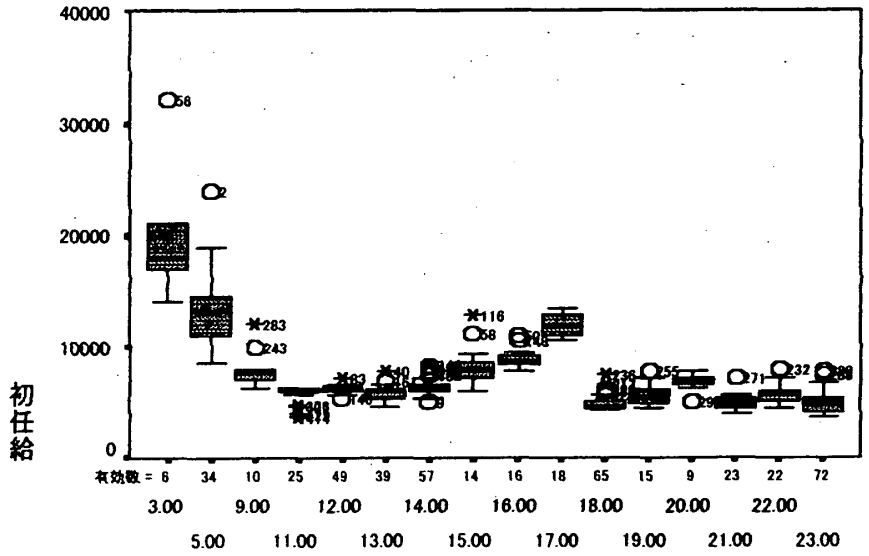
今回、停止則として、階層を3とし、親ノードのケース数を10、子ノードのケース数5を任意に選んだ。ノード数の根拠は何かしらの理論に基づいたものではなく、分析者の恣意による点が大きい。そうした恣意性を減じるために、「箱ヒゲ図と分散分析の事後分析(多重比較)を用いてターミナルノードの検証を行い、枝刈を行なう」ことを提案したい。

2−3　事後分析(多重比較)

分散分析後の多重比較において、同じ親を持つ子ノードで、5%で棄却されるターミナルノードの組み合わせは、13と14、14と15、13と15、16と17、18と19、18と20、21と22、22と23、である。この場合の帰無仮説は「平均値に差はない」ということであるので、これらのターミナルノードの組み合わせは1つのノードにまとめない方がよいようだ。

一方、5%で棄却されないターミナルノードの組み合わせは、11と12、19と20、21と23、である。したがって、これらのターミナルノードの組み合わせはそれぞれ1つのノードにまとめることが考えられる。

図2　ターミナルノードの層別箱ヒゲ図



以上のことから、ターミナルノード11と12の組み合わせは親ノード4にまとめることに疑問は残らない。しかし、子ノードが3個あり、そのうちの1組だけしか棄却されない18から20、21から23、の組み合わせはどう扱うかが難しい。今回はひとまずこれらを親ノードにまとめることにした。結局、ターミナルノードは16個から11個へと減らすこととなった。

3．まとめ

データ解析の取り付き難さは、統計の初心者であっても決定木分析を利用することで減ずることができた。しかし、枝刈に関して多重比較を用いることは、指導者のアドバイスによるものである。今後は、別の観点から停止則に関して検討を行ないたい。

参考文献

1　豊田秀樹(2001).『金鉱を掘り当てる統計学　データマイニング入門』, 講談社.

2　新村秀一(1995).『パソコンによるデータ解析』, 講談社.