決定木による判別と予測

奥　喜正　・　内桶　誠二

1．はじめに

経営情報戦略で有効な戦略ツールとなってきたデータマイニングにおいて，頻繁に利用されるデータ解析メソッドとして伝統的な統計的データ解析手法の回帰分析，クラスター分析とともに，データマイニング特有のメソッド（Data-adaptive Method）として発展してきた，決定木（Tree Models）やニューラルネットワークがある（Berry and Linoff, 2004）。データマイニングの実行目的はルールの発見にあるが，具体的には判別，予測，連関，分類諸問題であり（福田 他，2001），本稿では判別と予測問題への決定木の有効活用を模索する。

従来の判別と予測では，判別分析，線形回帰分析やロジスティック回帰分析を利用してきた。決定木の応用目的の一つに優良顧客（Loyal Customers）などの判別があるが（杉田・桜井，2001），決定木により生成される判別ルールを利用して大量の大規模データベースから優良顧客データのみを抽出することが高速に可能になる。そこで，本稿ではタイタニック号乗客生死データを改良したデータセットを利用して，優良顧客の判別ルールの作成過程についてCHAID（CHi-square Automatic Interaction Detection）によって最初に説明する。つぎに，CART（Classification and Regression Trees）の回帰決定木による目的変数の予測精度について回帰分析のそれと比べた場合の優劣をモデル検証法によって（Hjorth, 1994），住宅費データ（Harrison and Rubinfeld, 1978）を使用して検討する。説明変数の数が多いときに回帰分析で解析を行う際には難解なる変数選択問題が生ずるが，変数選択の一方法として決定木を利用して説明変数群を絞り込むことが有効であると考えられる。また，変数選択問題を回避して目的変数の予測に回帰決定木を直接的に利用することも可能である。そこで，回帰決定木による目的変数の予測精度を伝統的なデータを利用して検討し，さらに，回帰決定木が作成する明示的な木構造によるルールを例証することで，回帰分析の変数選択方法よりもユーザにとって理解しやすい説明変数選択を回帰決定木が成せる可能性を示唆する。

本稿では決定木の作成する判別ルールによる特定データの抽出方法を示し，つぎに，回帰決定木で予測した場合の予測精度をクロスバリデーション（Cross Validation）というモデル検証法で検討し，判別・予測問題に関して決定木というノンパラメトリックモデル（Non-parametric Predictive Model）を適用することの有望性を検討する。

2－1．決定木について

決定木は一連の説明変数の中から適切な変数を選択して，順次データセットを，より均質なサブグループに分割して目的変数の予測や判別を行う手法である。モデルに対して正規分布など確率分布の仮定が不要なノンパラメトリック手法で，多くの説明変数が扱えるなどの利点が

決定木にはある（Giudici, 2003）。

決定木手法としてはCHAID（CHi-square Automatic Interaction Detection）が代表的である（Kass, 1980）。また，昨今，広く利用されている手法にCART（Classification and Regression Trees）がある（Breiman et al, 1984）。一般的に目的変数が質的な場合にはCHAID，CARTを利用し，量的変数のときはCHAIDの使用実績が少ない為にCARTを使用する傾向が多い（SPSS, 2001）。

決定木では一連の説明変数の中から，一つの適切な説明変数を選択してデータセットを，より均質な傾向をもつサブセットに分割することを繰り返して，目的変数に強く関連している説明変数や注目したいサブグループを発見することを目的とする手法である。クラスター分析では，データ分割のためにすべての変数を利用するのに対して，決定木ではデータ分割は一つの説明変数のみで行われる点が決定的に異なる。

2－2．CHAID

CHAIDによる解析は，つぎのような手順を再帰的に繰り返して木を作成することである。まず，目的変数と各説明変数間でクロス集計表を作成する。つぎに，クロス集計表にカイ二乗検定を実行して，最も有意である，すなわち，P-valueが最小の説明変数をデータ分割のための説明変数の候補とみなしてそれによってデータセットを分割する。

CHAIDはクロス集計表に対する独立性のためのカイ二乗検定を行い，そのP-value値をデータセット分割規準に使用して，繰り返しデータセット分割を停止条件が成立するまで実行して決定木をつくるアルゴリズムである。決定木の評価には，CHAIDでは不純度の測度（Impurity Measure）としてピアソンのχ^2検定統計量を使用する。また，分割対象データが全データの１％というのが，それ以上はデータセット分割を行わない分割停止条件の目安の一つである。

また，有意水準をαと設定した検定をｎ回繰り返すと，第１種の誤りを犯す確率は$1-(1-\alpha)^n \fallingdotseq n\alpha$となる。有意水準５％の検定を４回実行すると，検定全体の有意水準は18.5％になってしまうので，有意水準を検定回数nで割るというBonferroni の調整済み確率をCHAIDでは採用して，P-valueを総括的に調整する。クロス集計表では説明変数の数が多いときに，その組み合わせ数が多くなって集計表の分析結果が煩雑になってデータ構造の解読が困難になる。このような時にCHAIDは解析結果を明示的な木構造にまとめて，視覚的に理解しやすいルールを提供する。

2－3．CART

CART（Classification and Regression Trees）は目的変数が量的な場合の応用でCHAIDに比べてその歴史が古い。ここで，CARTのアルゴリズムについて言及する。このアルゴリズムはGini Impurityを不純度の測度として，それを最適にするようにデータセットの二分割を行い，かつ，枝刈り（Concept of Pruning）を行う。すなわち，決定木を過学習状態（Overfitting）になるまで十分大きく成長させてから，過学習に相当するリーフやノードの部分木を削除して枝を刈り，最善のサブツリーを発見しようとするのが，CARTアルゴリズムの基本方針である。ここで，過学習とは学習用データの例外的な値などに過度に適合する枝葉のことで，一般的なデータに対しては必要以上に細かい分類になっていて，かえって予測精度を悪くしてしまう状態をいう。枝刈りの方法には十分に大きな木を生成してから過学習に相当する部分木を取り除く「事後枝刈り（Post-pruning）」の手法が現在は多く利用されている（福田 他, 2001）。事後枝刈り方法には，検証データを使用して部分木を削除する手法が特に利用されており，中でも複雑度の増加を考慮するCost-complexity Pruning手法が現在，標準的な方法になっている（Breiman et al., 1984）。CARTアルゴリズムの特徴は，データセットを常に二分岐させて成長させること，分割において統計的検定を利用

しないこと，木を過学習状態になるまで成長させてから枝刈りを行う点に要約されよう。

CARTは，目的変数の種類により二分別できる。目的変数が質的変数のときは分類樹木（Classification Trees）となり，不純度はGini測度で測定されて不純度を最も減少させる説明変数でデータセットを二分割する。なお，Gini測度は，$1-\sum_{i=1} p_i^2$（ここで，p_iは比率）で示される。一方，目的変数が量的変数のときは回帰樹木（Regression Trees）となり，この場合の不純度はノード内分散で測定されて，ノード内分散の最小化を目標として回帰決定木は作成される。

非線形性の可能性が大きいときや交互作用の可能性がある場合には，線形回帰分析の使用に先立って回帰樹木を利用すべきであるという見解もある（SPSS, 2001）。一般的に，決定木では過学習を防ぐためにもモデル検証を行うべきで，学習用データを全データの50～90%，検証用データは10%～50%とすることが多い。

3．優良顧客判別のためのCHAIDの実行

タイタニック号乗客生死データに改良を加えた2101名から成る，改良データセットにSPSSのAnswerTreeのCHAIDを適用しながら，決定木形成過程を説明する。元来のデータは生存，死亡という目的変数であるが，目的変数の値をCRM（Customer Relationship Management）的なものに置き換え，生存・死亡を優良顧客・不良顧客と読み替えて（奥 他, 2004），経営情報学におけるCHAIDの有効活用を説明する。本稿のデータセットでは目的変数は「乗船客の優良顧客，不良顧客」で質的変数である。説明変数には「性別」，「(大人か子供かという）年齢」，乗船した「等級」が選ばれている。

本稿のCHAIDによる決定木は，つぎのような条件で形成された。すなわち，データセットの分割基準（Division Criteria）には，Bonferroniの調整済み確率によるp-valueを採用し，停止条件は最小ノード数（Mnimum Node Size）が10，決定木の深さ（Maximun Tree Depth）は3に設定した。

最初の分割ステップでは，乗船客優良・不良という目的変数と，性別，年齢，等級という説明変数間で3個のクロス集計表を作成して独立性のためのカイ二乗検定を実行してそれぞれのP-valueを求める（表1）。一番小さいP-valueの説明変数をデータセット分割のための基準変数とするため，説明変数「性別」がこの分割ステップの基準変数に選ばれた。

表1　最初のステップのデータ分割

	カイ二乗値	自由度	P-value
性別	447.2	1	2.83E-99
等級	187.9	3	1.76E-40
年齢	17.18	1	3.39E-05

つぎの分割ステップに進む。男性では，男性と年齢，等級の二個のクロス集計表が作成可能である。

表2　ステップ2（男性）

男性		自由度	カイ二乗値	p-value
	年齢	1	22.3	2.33E-06
	等級	3	29.05	5.72E-06

男性サブセット分割では，表2のP-Valueの値から分割のための説明変数には「年齢」が採用された。女性サブセットの第二分割では，分割の説明変数にはp-valueの値から「等級」のみが5%有意な変数として採用された。なお，「年齢」のP-Valueは0.027（$>0.05/2=0.025$）であるので，5%有意な説明変数とはいえない。

表3　ステップ2（女性）

女性		自由度	カイ二乗値	p-value
	年齢	1	4.845	0.027726395
	等級	3	125.28	5.62472E-27

男性サブセットのさらなる分割では，「大人」，「子供」の層で共に「等級」が，分割のための説明変数として採用された（図1）。ここで着目すべきことは，二等乗船客より三等乗船客のほうが成人男性では優良率が高かったとい

う事実である。また，男の子供の一，二等乗船客の優良率は，100%であった点も興味深い。

決定木の具体的な応用では，顧客データベースからの優良顧客の属性モデル化を挙げることができる。ここでは，図2から「女性一等，二等乗船客，かつ，男の子供の一等，二等乗船客

男性		カイ二 乗値		P-Value
	子供	24.027	2	0.000
	大人	36.924	3	0.000

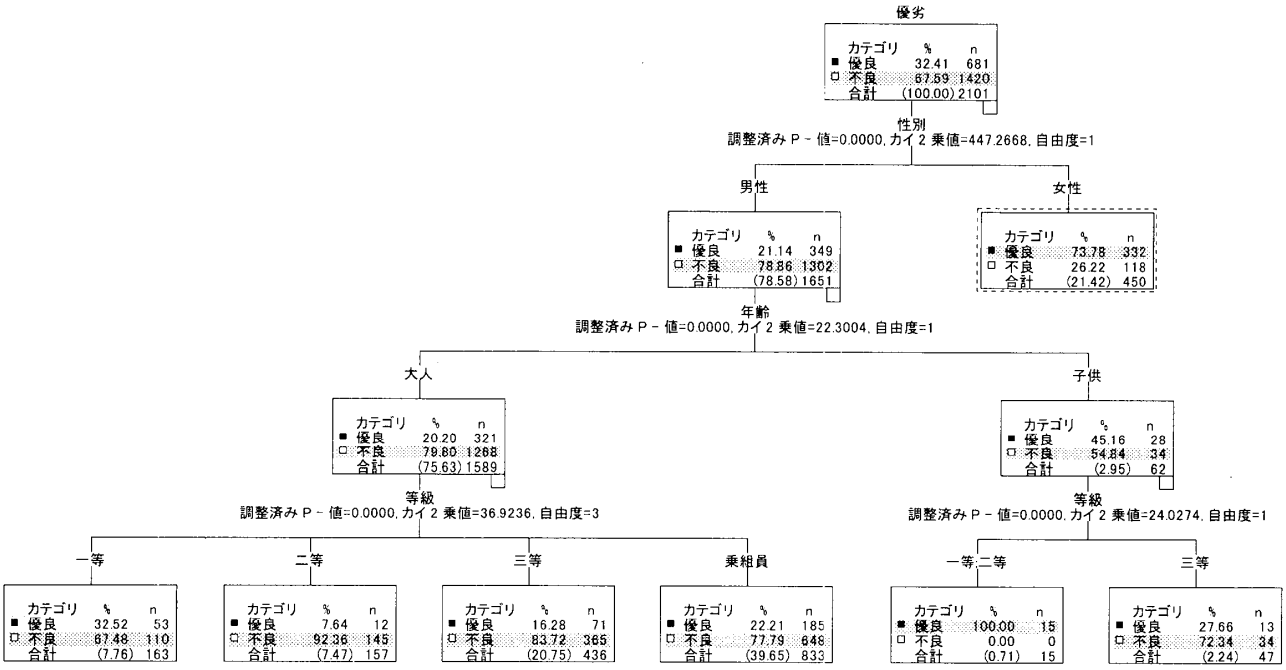

図1　ステップ3（男性の部分木）

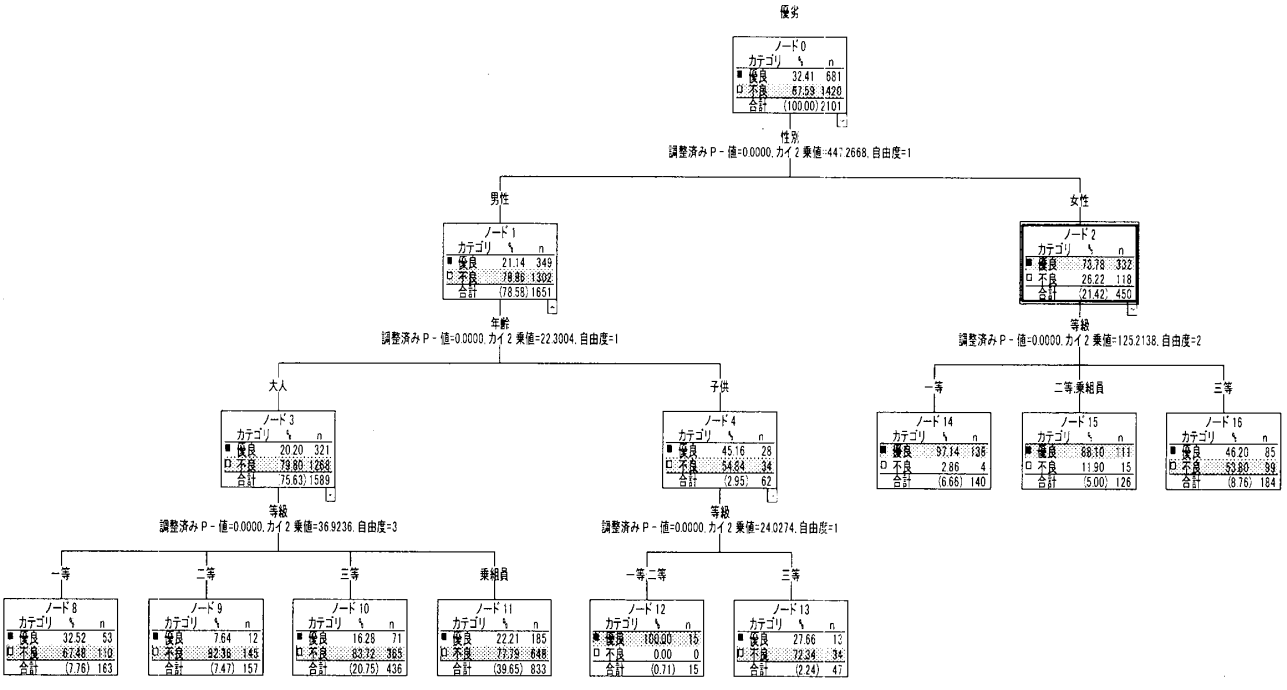

図2．最終的なCHAIDによる決定木

を優良顧客とみなす」という判別ルールが生成されて，それをSPSSのシンタックス言語でつぎのように示す。

```
* ノード12，5，6．
SELECT IF (((MISSING(性別) OR 性別 NE 0) AND (年齢 EQ 0) AND (等級 EQ 1 OR 等級 EQ 2) OR (性別 EQ 0) AND ((等級 EQ 1) OR (等級 EQ 2 OR 等級 EQ 0)))) .
EXECUTE.
```

これをSPSSのシンタックス言語として，全データから上記条件を満たす顧客データのみを選択的に優良顧客として抽出することが高速に可能になる。

4．回帰決定木の予測精度

説明変数から目的変数の値を予測する場合，従来の統計的データ解析では回帰分析を用いてきた。説明変数の数が多いときには，ステップワイズ法などの変数選択法を当該モデルでは使用することが多いが，現在でも変数選択問題は統計学の難しい課題になっている。回帰決定木を使用して説明変数を絞り込むことも昨今は提案されているが，説明変数の数が多いときに，直接的にCARTの回帰決定木によって目的変数を予測した場合に，線形回帰分析のそれと比較して予測精度が如何なるものになるかをつぎに検討する。

統計的データ解析ツールとしての線形回帰分析モデルは次にように表記される。

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\varepsilon \qquad \varepsilon\sim N(0,\ \sigma^2)$$

誤差項εへの仮定を明確にするためにベクトル表示すると，標準的な線形回帰モデルの誤差ベクトル$\boldsymbol{\varepsilon}$は

$\boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\ \sigma^2 I)$　（ここで，Ｉは単位行列）

という多変量正規分布に従うことになる。標準的線形回帰モデルは誤差項εに所定の条件を備えた正規分布を仮定するというパラメトリックな統計モデルである。この仮定によってパラメータ$\boldsymbol{\beta}$に関する検定などが実行可能になり，また，最小二乗推定量$\boldsymbol{\beta}$が最小分散不偏推定量であることが保証される。一方，決定木（Tree-structured Methods as Data Mining Tools）は，教師あり学習（Supervised Learning）の範疇に入る手法で，再帰的な分割を行うデータに対する確率分布の仮定を要求しないノンパラメトリックな手法である。

さて，回帰分析と回帰樹木の予測精度をクロスバリデーション(Cross Validation)を利用してモデル検証(Validation)を行う(Hjorth, 1994)。データマイニングの適用場面ではデータ量が一般に大きいので，所与のデータセットをモデル作成のための学習用データ(Training Data)と，作成されたモデルの精度を検討するために使用する検証用データ(Validation Data)に分けて利用することができる。全データを使用して作成した予測モデルでは，その誤差分散σ^2を過少評価してしまう恐れが大きいのでクロスバリデーションを行うことがある（Everitt, 2002)。

本稿で利用するデータは住宅費データ（Harrison and Rubinfeld , 1978）で，その論文では回帰モデルが使用されているが，目的変数及び説明変数群の具体的な説明を表４に示す。線形回帰分析はＦ値＝2.0のステップワイズ法の変数選択法で実行した。他方，回帰決定木は，CART により作成して停止条件は，樹木の深さは５，最小ノード数５，不純度の最小変化量を0.1に設定した。

全データによって作成した場合の線形回帰分析結果を表６に，他方，回帰決定木を図３に，樹木によって作成される目的変数予測のルールを付表に示す。付表のように回帰決定木による目的変数の予測値は，階層的条件をみたすターミナルノードのデータ平均値により推定される。すなわち，該当するターミナルノードに至る経路を示す階層的な分岐条件が目的変数値の予測ルールになる。

最初に，全データを使用した場合の線形回帰モデルおよび回帰決定木の残差二乗和$\sum(y-\hat{y})^2$及び，残差二乗の標準偏差を求めて，それぞれの予測精度を測定したところ表５のような結果になった。全データの場合は予測精度は線形回帰分析モデルよりも回帰樹木のほうが，残差二

乗和，その標準偏差の観点から高いという結果が得られた（表５）。

また，図３から住宅費が高いグループは，「住居あたりの部屋数の平均」が7.4より大きい場合と，あるいは，人口あたりの下層階級の％が14.4％未満で職業センター（Employment Centers）に近い地域であることが読み取れた。住居あたりの部屋数の平均が7.4より大きい層は，住宅費が45.1であるのに対して，住居あたりの部屋数の平均が6.9より大で7.4未満の場合は住宅費が32.1であるように，この説明変数の7.4前後の値の変化は重要な意味を持つことまで推測できた。住宅費の低いグループは，人口あたりの下層階級率が14.4％より大きくて犯罪発生率が6.9％より高い地域で，窒素酸化物濃度０.605以上の地域が最も安く（平均住宅費11.08），窒素酸化物濃度0.605未満の地域でさえ平均住宅費が16.6と低いことが回帰決定木によって理解できた。加えて，二回以上データの分割に利用された重要なる説明変数は「住居あたりの部屋数の平均」（３回），「人口当たりの下層階級の％」（２回）であった。続いて「職業センターへの距離」，「犯罪発生率」が，決定木の分岐条件に使用され重要な説明変数であることが当該決定木から読み取れた（図３）。

つぎに，回帰分析，回帰決定木それぞれの予測精度を吟味するためにクロスバリデーション（Cross Validation）を実行した。すなわち，一様乱数を利用してデータセットを２分割し，一方を学習用データセット(50％)，他方を検証用データセット(50％)としてモデル検証に使用する。学習用データセットによって回帰分析の予測式，回帰樹木による予測ルールを作成した。つぎに，学習用データセットで作成した回帰予測式およびCARTによる予測ルールを検証データにあてはめて，それぞれの予測精度を測定した。このようなモデル検証を10回実行した結果を表７にまとめた。解析結果から全データを使用した場合には，回帰樹木の予測精度の方が優れていたにもかかわらず（表５），クロスバリ

表４　住宅費データの変数一覧

	Explanations of Respective Variables
説明変数１	Crime Rate by Town
2	Proportion of a Town's Residental Land Zoned for Lots Greater Than 25000 Square Feet
3	Proportion of Nonretail Business Acres per Town
4	Chales River Dummy
5	Nitrogen Oxide Concentrations in Pphm
6	Average Number of Rooms in Owner Units
7	Proportion of Owener Units built Prior to 1940
8	Distances to Five Employment Centers in the Boston Region
9	Index of Accessibility to Radial Highways
10	Full Value Property Tax Rate
11	Pupil-teacher Ratio by Town School District
12	Black Proportion of Population
13	Proportion of Population that is Lower Status
目的変数	Median Value of Ower-occupied Homes

表５　回帰分析と回帰決定木の予測精度（全データ使用の場合）

	残差二乗和	標準偏差
回帰分析	11,081.36	59.169
回帰樹木	7,904.87	44.319

表６ 回帰分析の有意な説明変数一覧(全データ使用)

係数a

モデル		非標準化係数		標準化係数	t	有意確率
		B	標準誤差	ベータ		
11	(定数)	36.341	5.067		7.171	.000
	人口あたりの下層階級の%	-.523	.047	-.406	-11.019	.000
	住居あたりの部屋数の平均	3.802	.406	.290	9.356	.000
	その町の生徒と教師の割合	-.947	.129	-.223	-7.334	.000
	5つあるBoston職業センターへの距離	-1.493	.186	-.342	-8.037	.000
	pp 10M中の窒素酸化物の濃度	-17.376	3.535	-.219	-4.915	.000
	Charles川からの距離	2.719	.854	.075	3.183	.002
	町単位での黒人の割合	.009	.003	.092	3.475	.001
	25K以上の居住面積の割合	.046	.014	.116	3.390	.001
	犯罪発生率	-.108	.033	-.101	-3.307	.001
	高速道路へのアクセス回数	.300	.063	.284	4.726	.000
	総資産税($10万)	-.012	.003	-.216	-3.493	.001

a. 従属変数: 持ち家の住宅費の中央値(1000単位)

表７　モデル検証による回帰樹木と回帰分析の予測精度比較

	回帰樹木		回帰分析	
検証回数	残差二乗和	標準偏差	残差二乗和	標準偏差
1	5276.98	61.42	6088	71.5
2	7043.43	99.77	5161.16	59.18
3	5062.4	53.7	5941.11	47.89
4	4226.34	48.57	4959.41	62.92
5	6658.17	84.43	5169.97	60.13
6	5588.05	72.12	6695.38	72.85
7	5062.44	73.74	4639.74	37.68
8	3176.69	24.35	5562.95	39.24
9	5062.4	53.7	5941.11	47.9
10	7043.43	99.77	6274.61	63.28

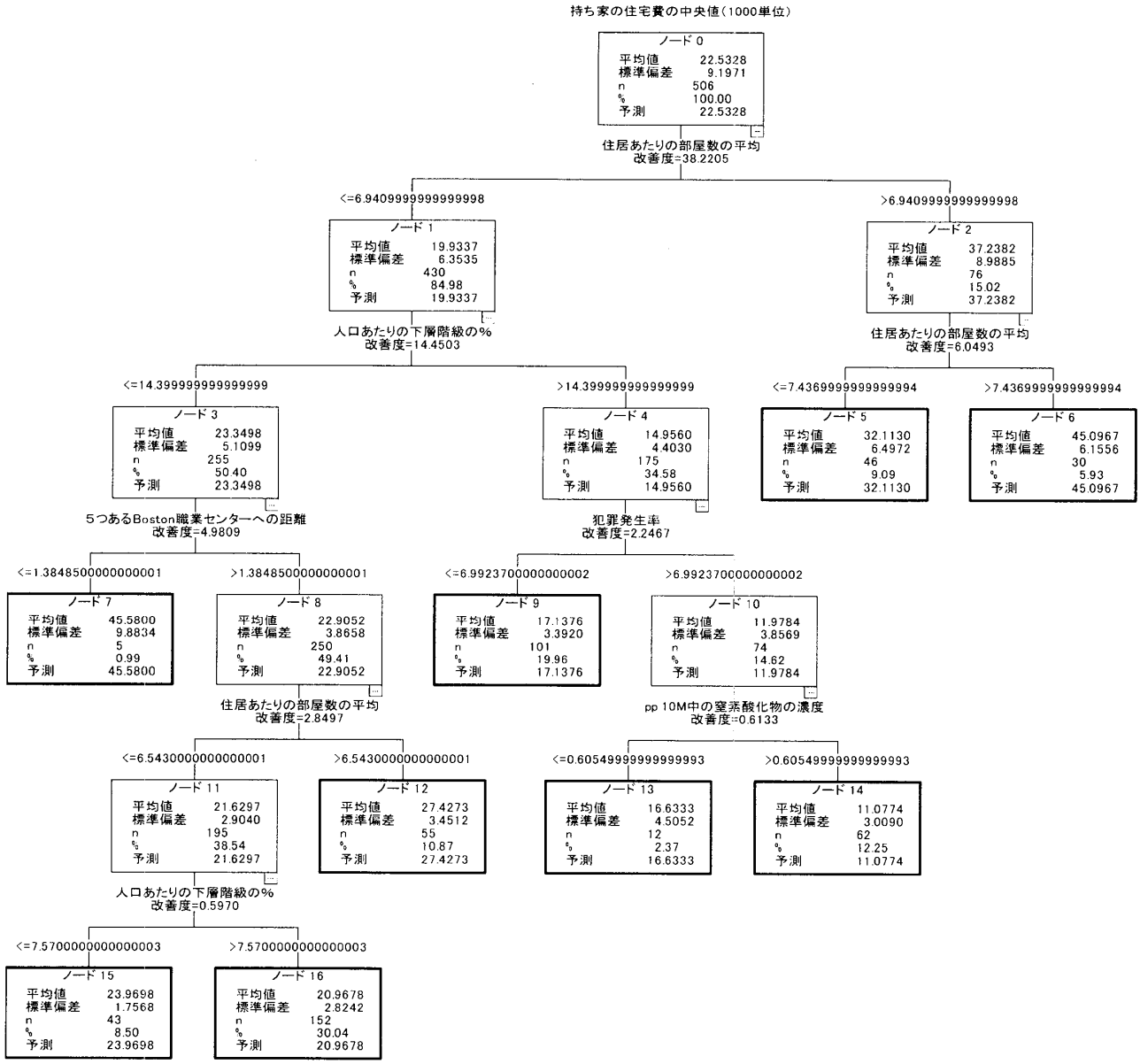

図３．住宅価格データの回帰樹木（全データ使用の場合）

デーションの場合には回帰樹木が優れている場合が４ケース，他方，回帰分析が良好な場合が４ケースというように結果合い半ばという意味深長なものになった(表７)。

５．考察

住宅費を利用した本稿のデータ解析では，全データでは回帰決定木の予測精度が優れたいたのにモデル検証結果では決定木の予測精度が回帰分析に比較して必ずしも良好ではないという事実が判明し，クロスバリデーションによって本稿の回帰樹木が学習データに過度に適合しているという，過学習状態にあることが検出され

たとも見做せる。

つぎに，線形回帰分析における誤差分散σ^2の不偏推定値を求める。

表８ 誤差分散σ^2の不偏推定値

	誤差分散の推定値
全データの場合	22.432
検証データの場合	23.416

誤差分散σ^2の不偏推定量は次式で求められる。

$$\hat{\sigma}^2 = \frac{\text{残差二乗和}}{\text{データ数}-\text{未知パラメーター数}} = \frac{\sum (y-\hat{y})^2}{(n-[p+1])}$$

モデル検証の場合の不偏推定値の平均は23.4で，全データの場合の誤差分散の不偏推定値22.4より大きいことが判明し，全データ利用では回帰分析での誤差分散の過小評価が示された。モデル検証の実行で，回帰分析での予測誤差σ^2の過小評価や決定木の過学習状態の発見が可能になったことは意義深い。このように，データマイニングではデータ量が大きいためにデータセットを二分割してモデル検証が実行可能になったことが，データマイニングと従来の統計的データ解析との重要な相違点といえるかもしれない。

回帰樹木は回帰分析の代用メソッドとして注目されている。回帰決定木を実行して，重要な説明変数の抽出や，交互作用や非線形関係を特定化してから，従来の回帰分析を実行すべきであるという考え方もある。回帰樹木は，説明変数が多い場合，はずれ値や異常な分布，高次の交互作用の可能性がある場合には回帰樹木を使用すべきであるともいわれている(SPSS, 2001)。また，回帰分析では目的変数にとって重要な説明変数は，有意であるか否かで判断されて，多数の説明変数間の階層構造や重要性の序列までは明示できないが，回帰決定木では本稿の例のように最初のデータ分割に利用された説明変数，すなわち「住居あたりの部屋数の平均」が最初の分岐変数，すなわち，最重要な変数であることなどが視覚的に理解できる(図３)。

回帰決定木では，諸説明変数の階層構造関係及び，重要な説明変数の詳細なる説明が付加される。かつ，予測力も回帰分析と同程度であることを鑑みると将来的に回帰決定木が有効ツールとなることが予想できる。また，判別を目的とした決定木の判別ルールにより優良顧客データの抽出が効率的に実行できることも本稿で示唆した。このように，判別分析や回帰分析の代替方法として判別・予測問題に決定木を適用することが極めて有効であることを本稿は示唆した。

参考文献

Berry, M and Linoff, G. (2004). Data Mining Techniques for Marketing, Sales, and Customer Relationship Management, 2nd ed. Wiley, New York.

Breiman,L., Friedman J.H., Olshen,R.A. and Stone,C.J. (1984). Classification and Regression Trees. Belmont, Calif, Wadsworth.

Everitt, B.S. (2002). The Cambridge Dictionary of Statistics, 2nd ed. Cambridge University Press, Cambridge.

Giudici, P. (2003). Applied Data Mining: Statistical Methods for Business and Industry. Wiley,New York.

Harrison,D. and Rubinfeld D.L. (1978). Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5, 81-102.

Hjorth,J.S.U.(1994).Computer Intensive Statistical Method. Chapman & Hall,London.

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29,119-127.

Miller,T.W. (2005). Data and Text Mining. Pearson Prentice Hall, New Jersey.

SPSS (2001). AnswerTree 3.0J User's Guide, SPSS Inc.

朝野 熙彦 (1998). 消費者行動の予測を目的としたマーケティング・セグメンテーション. マーケティングサイエンス, 6, 45-66.

SPSS著, 杉田善弘・桜井聡訳. (2001). マーケティングのためのデータマイニング入門 東洋経済新報社.

奥 喜正・本村猛能・前鶴政和・内桶誠二 (2004). データマイニングにおける二値データ解析: 決定木とロジスティック回帰分析. 物流問題研究, 44, 1-14.

佐和隆光 (1979). 回帰分析. 朝倉商店.

福田 剛志・森本 康彦・徳山 毅 (2001). データマイニング. 共立出版.

付表. 全データ使用の場合の回帰決定木による予測ルール (SPSS シンタックス)

```
ノード 7.
DO IF (SYSMIS(RM) OR (VALUE(RM) LE 6.941)) AND (SYSMIS(LSTAT) OR (VALUE(LSTAT)LE 14.4)) AND (DIS LE 1.38485).
        COMPUTE nod_001 = 7.
        COMPUTE pre_001 = 45.58.
END IF.
EXECUTE.
* ノード 15.
DO IF (SYSMIS(RM) OR (VALUE(RM) LE 6.941)) AND (SYSMIS(LSTAT) OR (VALUE(LSTAT)LE 14.4)) AND (SYSMIS(DIS) OR (VALUE(DIS) GT 1.38485)) AND (SYSMIS(RM) OR
(VALUE(RM) LE 6.543)) AND (LSTAT LE 7.57).
        COMPUTE nod_001 = 15.
        COMPUTE pre_001 = 23.969767.
END IF.
EXECUTE.
* ノード 16.
DO IF (SYSMIS(RM) OR (VALUE(RM) LE 6.941)) AND (SYSMIS(LSTAT) OR (VALUE(LSTAT)LE 14.4)) AND (SYSMIS(DIS) OR (VALUE(DIS) GT 1.38485)) AND (SYSMIS(RM) OR (VALUE(RM) LE 6.543)) AND (SYSMIS(LSTAT) OR (VALUE(LSTAT) GT 7.57)).
        COMPUTE nod_001 = 16.
        COMPUTE pre_001 = 20.967763.
END IF.
EXECUTE.
* ノード 12.
DO IF (SYSMIS(RM) OR (VALUE(RM) LE 6.941)) AND (SYSMIS(LSTAT) OR (VALUE(LSTAT)LE 14.4)) AND (SYSMIS(DIS) OR (VALUE(DIS) GT 1.38485)) AND (RM GT 6.543).
        COMPUTE nod_001 = 12.
        COMPUTE pre_001 = 27.427273.
END IF.
EXECUTE.
* ノード 9.
DO IF (SYSMIS(RM) OR (VALUE(RM) LE 6.941)) AND (LSTAT GT 14.4) AND (SYSMIS(CRIM) OR (VALUE(CRIM) LE 6.99237)).
        COMPUTE nod_001 = 9.
        COMPUTE pre_001 = 17.137624.
END IF.
EXECUTE.
* ノード 13.
DO IF (SYSMIS(RM) OR (VALUE(RM) LE 6.941)) AND (LSTAT GT 14.4) AND (CRIM GT 6.99237) AND (NOX LE 0.6055).
        COMPUTE nod_001 = 13.
        COMPUTE pre_001 = 16.633333.
END IF.
EXECUTE.
* ノード 14.
DO IF (SYSMIS(RM) OR (VALUE(RM) LE 6.941)) AND (LSTAT GT 14.4) AND (CRIM GT 6.99237) AND (SYSMIS(NOX) OR (VALUE(NOX) GT 0.6055)).
        COMPUTE nod_001 = 14.
```

```
        COMPUTE pre_001 = 11.077419.
END IF.
EXECUTE.
*ノード 5.
DO IF (RM GT 6.941)  AND  (SYSMIS(RM) OR
(VALUE(RM) LE 7.437)).
        COMPUTE nod_001 = 5.
        COMPUTE pre_001 = 32.113043.
END IF.
EXECUTE.
*ノード  6.
DO IF (RM GT 6.941)  AND  (RM GT 7.437).
        COMPUTE nod_001 = 6.
        COMPUTE pre_001 = 45.096667.
END IF.
EXECUTE.
```