ハイブリッドモデリングに関する研究

クオリティマネジメント研究　699B016-0 小方英樹
指導 棟近雅彦 教授

A Study on Hybrid Modeling

By Hideki Ogata

1.研究背景および目的

ハイブリッドモデリング(以下 HM)とは，ノンパラメトリック手法の樹形モデルの一つである Classification and Regression Trees(以下CART)と従来の統計手法である回帰，判別，ロジスティック回帰分析を組み合わせたモデルである．

HM は，Steinberg[1]らによって提案され，大滝ら[2][3][4][5]によって改良された．それにより，欠測値のあるデータへの適用や，データに内在する新たな情報の抽出に有効であることがわかった．しかし，大規模データでの適用例がなく，また HM の作成手順が確立したとはいえない．

そこで本研究では，中古車のオークション情報という大規模データを用い，HM の有効性の検証と手法の確立を目的とする．また同時に CART の方法についても言及する．

2.従来の研究と問題点

2.1 大滝らが提案した HM

大滝らは，実際に HM を適用可能にするために，Steinberg らの提案を参考に，数例の解析経験に基づいて検討を行った．そして以下のような HM の具体的な手順を提案した．

［ステップ 1］CART による 2 進木の選択を行う．
ただし，以下の規則で行う．

- CART はデータ変換をせずに原データによる樹形分析を行う．
- 剪定基準として，標準誤差を利用する．
- 誤差の検証法として，交差検証法を利用する．
- ターミナルノードをダミー変数化する．

［ステップ 2］パラメトリックモデルに CART の結果を適用する．

- 説明変数候補としてダミー変数化した CART のターミナルノード，変数重要度*1 の高い変数，最初の分岐で競合変数*2 となったものをあげる．
- 欠損値に対しては，代理変数*3 を利用する．

＊1 すべての分岐変数と代理変数の改善度の合計に対するある変数の改善度の割合
＊2 分岐の際に分岐変数と競合する変数
＊3 最適な分岐変数が欠損値の場合に用いる変数

［ステップ 3］時系列データへの対応を行う際，パラメトリックモデルの残差部分に自己相関モデルを適用する．

2.2 従来の研究の問題点

大滝らの方法を検討した結果，以下の問題点が挙がった．

- **データ数**

従来の研究において適用されたデータ数は，せいぜい 1000 程度であった．そこで本研究においては，中古車価格のデータに適用することにより，一万個以上のデータを扱い，大量データの解析を行い，新たな知見を得る．

- **CART の方法**

従来の研究においては，HM における CART の方法があいまいであった．例えば改善度*4 が同じ場合の対処方法や，停止規則に関する考察が不十分であった．そこで本研究においては，HM における CART の方法を検討し確立する．なお，CART のプログラムは SPSS の answer tree2.0 を用いた．

＊4 分岐したときの残差平方和の改善度

- **HM の変数候補**

従来の研究においては，明確な理由を示さずに，HM の変数候補がある程度限定されていた．そこで本研究においては，説明変数すべてを変数候補とする方法と，大滝らが提案した方法とを比較し，どちらの方法が優れているのか検証を行う．

3.本研究の提案

前述した問題点を考慮して，HM の方法の提案と検討内容を示す．

(1).データ変換はせずに原データによる樹形分析(CART 分析)を行う．ただし以下の規則で行う．

分岐規則：原則として改善度が最大である変数で分岐する．ただし，改善度が等しかった場合については，今までの知見から重要であると判断できる変数，あるいはそれぞれ選択し良い結果が得られる方の変数を選択する．

停止規則：以下の基準のいずれかに当てはまった時点で停止する．

- 樹木の最大の深さ(階層数)に到達する．
- ノードを分岐させる有意な予測変数が存在しない．
- ターミナルノード内のケース数が，親ノードの最小ケース数を下回る．
- 仮にノードを分岐させた場合，一つまたは複数の子ノード内のケース数が，子ノード用の最小ケース数を下回る．

剪定規則：標準誤差と最小誤差を比較する．

誤差検証法：すべての検証法を比較する．

- 代替推定法：すべてのデータを学習用と検証用に用いる．
- テストサンプル法：データを学習用と検証用に分ける．
- 交差検証法：データを検証用として任意のグループに分け，その誤差推定値を平均する．一般的にデータの少ないときに使用する．

(2).個々のケースのターミナルノードをダミー変数としてパラメトリックモデルに導入し，分析を行う．ただし変数は以下のように扱う．

変数候補：CART の結果を利用したダミー変数を含むすべての変数あるいは大滝等の方法による変数とする．

説明変数の選択：原則として変数増減法で行い，共線性が生じた場合には過去の知見から判断する．

欠損値：代理変数を利用する．

4.事例への適用

4.1　データの概要

本研究で利用したデータを以下に示した．本研究においては，中古車の価格が量的目的変数であることより，CART と回帰分析を組み合わせた HM を扱うこととする.中古車オークションとは，中古車販売会社や一般ユーザーが全国各地の会場で中古車を売買することである．

- データ数：22716 件
- 期間：1998 年 12 月～2000 年 5 月
- 目的変数：中古車オークションでの中古車の落札価格
- 説明変数：排気量，型式等の 26 項目

4.2　誤差検証法の比較

誤差検証法を比較するために，それぞれの検証法の推定誤差を表 1 に示した． ただし，交差検証法は代替推定法の最小誤差と結果が一致する．よって本研究においては交差検証法を省略し，代替推定法とテストサンプル法を適用した．

表 1：誤差検証法の比較(抜粋)

推定法	代替推定法	テストサンプル法 10%)		テストサンプル法 30%)		テストサンプル法 50%)	
		学習用	検証用	学習用	検証用	学習用	検証用
推定誤差	62374	70792	87527	68633	77796	73267	77355
標準推定誤差	1545	1601	7282	1784	3395	2109	2446
合計ノード数	69	35		39		31	
合計段階数	9	7		8		7	
合計ターミナルノード数	35	18		20		16	
剪定乗数	10標準誤差	10標準誤差		10標準誤差		10標準誤差	
推定法	代替推定法	テストサンプル法 10%)		テストサンプル法 30%)		テストサンプル法 50%)	
		学習用	検証用	学習用	検証用	学習用	検証用
推定誤差	60942	64448	82117	65756	75471	71420	75913
標準推定誤差	1542	1589	7275	1778	3387	2104	2443
合計ノード数	105	91		67		43	
合計段階数	10	10		10		7	
合計ターミナルノード数	53	46		34		22	
剪定乗数	最小誤差	最小誤差		最小誤差		最小誤差	

テストサンプル法間で比較すると，検証するサンプルが多いほど，標準推定誤差は低い値を示している．また，合計ターミナルノード数を比べると，標準誤差においては検証用サンプルとして全体の 30%をとる(以下検証 30%)方法，そして最小誤差においては検証 10%の方法が一番分岐していることがわかる．

次に代替推定法とテストサンプル法を比較すると，テストサンプル法(検証 30%)の最小誤差の結果とほとんど一致している．この代替推定法は学習用と検証用のデータが一致している推定法である．よって，新しいデータに対応できるかどうかは疑問であり，検証法を比較するためには，HM の重回帰分析の寄与率と残差標準偏差も考慮する必要がある．

4.3　標準誤差と最小誤差の比較

大滝らは HM を行う際に，標準誤差を利用している．そこで本研究においては，最小誤差と標準誤差両方で解析を行い，結果を比較した．

まず，最小誤差と標準誤差の違いについて考察する．標準誤差による剪定とは，なるべく小さく最小誤差に近い部分木を選択する方法である．よって，最小誤差を利用した樹木よりも，標準誤差を利用した樹木の方が合計ターミナルノード数は少なくなる．そのため一つのグループのデータ数が少なく，外れ値の影響が出やすくなる．

次に，分岐変数や分岐基準について細かく比較したところ，上位の分岐に関してはまったく同じ

であるが，分岐が下位になるほど分岐変数や分岐基準がわずかに異なっていることがわかった．

CART において多くの知見を得るためには，合計ターミナルノード数の多い最小誤差を剪定基準にすべきであると考えられる．しかし，確実な結果を得るためには，分岐が少なく外れ値に左右されない方法を行うことが重要である．つまり標準誤差を利用した樹木を利用した方が理にかなっていると考えられる．よって HM の一部として利用する際は，標準誤差を利用した方がいいと考えられる．

しかし最終的な判断は，4.2 と同様に，HM の重回帰分析の寄与率と残差標準偏差を比較し，精度を基準に決定すべきである．

4.4 HM の変数候補

HM の重回帰分析における説明変数と CART 結果のターミナルノードとの関連を検討する．ただし，本研究においては量的目的変数であることより，CART 以降の解析は重回帰分析を利用した．HM の重回帰分析で採用されなかった分類は，価格に大きな影響が見られなかったことを示している．その特徴を調べるために，散布図と改善度を指針に比較した．その結果，HM の重回帰分析で採用されている他の変数と違いを検出できなかった．また，HM の重回帰分析で採用される変数は，大滝らの提案したターミナルノード，変数重要度の高い変数，最初の分岐の競合変数だけでないこともわかった．

次に，精度を比較するため，重回帰分析による結果を表 2 に，大滝らの方法による結果を表 3 に，また本研究で提案した方法の結果を表 4 に示した．

表 2：重回帰分析の結果(抜粋)

比較項目	重回帰分析
寄与率	0.904
残差標準偏差	204.341

表 3：大滝らの方法による重回帰分析結果(抜粋)

比較項目	大滝らの方法
寄与率	0.916
残差標準偏差	204.341

表 4：HM の重回帰分析結果の比較(抜粋)

誤差検証法	テスト10	テスト30	テスト50	代替	剪定規則
寄与率	0.916	0.915	0.914	0.92	最小
残差標準偏差	203.845	205.763	206.55	199.83	
寄与率	0.917	0.916	0.914	0.92	標準
残差標準偏差	203.304	204.501	206.959	199.79	

表 3, 4 を比較すると，テストサンプル法の検証 10%と代替推定法の方が，寄与率と残差標準偏差から判断して，精度として優れていることがわかる．よって，HM の重回帰分析の変数候補は CART の結果を利用したダミー変数を含むすべての変数とする方がよいことがわかる．

4.5 重回帰分析と HM の比較

表 2, 4 を比較すると，寄与率と残差標準偏差から判断して，HM の重回帰分析の方が精度に優れていることがわかる．また 4.3 で示したように，CART によってわかりやすく確実な情報を抽出できることが分かる．よって，HM は重回帰分析よりも優れているといえる．

さらに，HM のモデルとして優れているのは，代替推定法の最小誤差と標準誤差となった．しかし 4.4 で示したように，最小誤差を採用することによって外れ値の影響が強く出ている可能性が強い．また CART の分岐基準やその値が一定しないと考えられる．さらに，CART の誤差検証法のテストサンプル法を利用した，HM の重回帰分析の精度を比較すると，学習用のデータが多いほどいい精度が出ることがわかる．このことより，大規模データにおいてテストデータを用意することは必要ないと考えられる．よって，HM の剪定規則は標準誤差を利用し，また誤差検証法は代替推定法を用いるべきであると考えられる．

5.最終的な提案

以上を踏まえ，最終的に HM の方法を提案する．ただし，3.の提案と異なる部分だけを以下に記述する．

(1).データ変換はせずに原データによる樹形分析 (CART 分析)を行う．ただし以下の規則で行う．

停止規則：以下の基準のいずれかに当てはまった時点で停止する．ただし，樹木の階層は最大まで成長できるようにする（6.1）．

- ノードを分岐させる有意な予測変数が存在しない．
- ターミナルノード内のケース数が，親ノードの最小ケース数を下回る．
- 仮にノードを分岐させた場合，一つまたは複数の子ノード内のケース数が，子ノード用の最小ケース数を下回る．

剪定規則：標準誤差を利用する(4.3)(4.5).
誤差検証法：代替推定法を利用する(4.2)(4.5).

(2).個々のケースのターミナルノードをダミー変

数としてパラメトリックモデルに導入し，分析を行う．ただし変数は以下のように扱う．

変数候補…CART の結果を利用したダミー変数を含むすべての変数とする(4.4)．

6.考察

6.1 分岐規則と停止規則

従来の研究においては，分岐規則，停止規則が明確ではなかった.これらの規則を明確にすれば，解析者が適用しやすくなる．そこで本研究では，3 節に示した規則を提案した.

分岐規則の原則としては，改善度が最大である変数で分岐するものとし，さらに改善度が等しい場合においても分岐規則を規定した．提案した方法により，一意的に CART での分岐方法を定めることができた．また，ターミネルノードが重回帰分析の変数に含まれることは妥当になったといえる.

停止規則の基準について考察する．CART の階層数によって改善度が変化するわけでなく，枝によっては改善度が大きな状態で分岐が停止することもありうる．その場合，階層数の停止規則によって停止した枝については，局所的な交互作用等の情報を把握できなくなる危険がある．したがって，枝を最後まで分岐させるために，樹木の深さを最大の階層数までと規定した方がよい.

6.2 CART の有効性

CART 分析における変数選択の際，従来の方法と同様に，過去の知見に基づいた解析も可能である．つまり，変数選択が自由に行うことができることにより，過去の知見を利用するだけでなく，隠れた情報を効率的に引き出せるという利点もあり，対話的な解析が可能である．よって，従来の方法の良い点が受け継がれているといえる.

次に CART は，剪定に標準誤差を利用することにより，外れ値によらない頑健性の高い分類ができる．また，結果が樹木上に表示できることにより，大規模なデータに対して有効な情報を発見しやすいという特長もある.

さらに，非線形部分の把握，質的変数のグループ化，交互作用の把握等，従来の解析においては検出しづらい情報を発見できる.

6.3 HM の有効性

4.5 で述べたように CART によって有効な情報を引き出し，代理変数の利用により欠損値に対応することが可能である．また，CART と同様に，HM の重回帰分析の変数選択を任意に指定できることにより，過去の知見を利用した解析あるいはモデル探索も可能である．さらには，HM と従来の重回帰分析を比較すると，精度の向上が顕著である.

6.4 本研究の意義

本研究においては，二万個を超える大量データを扱った．その結果，大滝らの方法と同様に従来の重回帰分析より精度が上がることがわかった．また，欠損値が多いデータや大量データにおいて有効な情報を，効率的に引き出せるという点で優れていることがわかった．また，本研究で提案した方法は，大滝らの方法よりも寄与率や残差標準偏差の点で優れている．さらに手順や規則を明確にしたので，適用が容易になった.

7.結論と今後の展望

本研究においては，HM における CART の方法を確定し，HM の方法を確立した．また，大規模データの解析において，HM は従来の重回帰分析や大滝らの方法による HM より優れているという結論に達した．今後は本研究で扱っていない質的目的変数の大規模データにおいても，本研究の方法が有効であるか検証する必要がある．また HM と MARS やニューラルネットワーク等の他のデータマイニング手法との比較を行うことが重要であると考えている.

参考文献

[1] Steinberg D. ,Cardell,N. S. (1998) :“ The Hybrid CART-LOGIT Model in Classification and Data Mining ”,the Eighth Annual Advanced Research Techniques Forum,American Marketing Association

[2] 門脇哲男・大滝厚(1999) :“ ハイブリッドモデリング(1) ”,JSQC 第 61 回研究発表会要旨集

[3] 鈴木教郎・大滝厚(1999) :“ ハイブリッドモデリング(2) ”,JSQC 第 61 回研究発表会要旨集

[4] 門脇哲男・鈴木教郎・大滝厚・鈴木督久(1999) :“ ハイブリッドモデリング(3) ”,JSQC 第 29 回年次大会研究発表会要旨集

[5] 門脇哲男・鈴木教郎・大滝厚・鈴木督久(1999) :“ POS データへのハイブリッドモデリングの適用 ”,JSQC 第 30 回年次大会研究発表会要旨集