実データに見る O2O 効果のいち事例

NRIマーケティング分析コンテスト2013

DATAEXPLORING 出口慎二

はじめに

- 本稿では，与えられたデータをもとに，AID（automatic interaction detector）によって，O2O（online to offline）効果の実例を見てみたい．
- onlineでの活動（たとえば情報収集）がofflineでの行動（とくに購買）に影響を及ぼすという指摘は，いまに限ったことでなく，以前より指摘がある．そうしたデータ（とくにonlineについて）をより得やすい環境は，近年のほうが，より整ってきている．そのため，かならずしも，そのすべてが常に有用なのか，どう使ったらよいのか分からない場合でも，大量のデータが日々蓄えられている．
- AID自体は1960年代からあるが，日本でも，おおむね2000年になるまでには，ひととおりの検討が重ねられ（たとえば日本分類学会），近年ではあまり注目を浴びてはいない感がある．
- 本稿では，とくに難しいことを考えず，与えられたデータを，特に加工もせず，そのままAIDで処理することで，自動的（automatic）に，どこに着目する価値があるか，それにはどの程度の期待が持てるか，の情報を得られることを，実例をとおして示したい．
- また，この目的において，やや不自然な手法上の制約（独立変数を繰り返し分岐変数として用いない）が，AIDの扱いやすさを増す可能性があることも，実例をとおして示したい．

■ AID（Automatic Interaction Detector）について

- ◆ Morgan, J. N. & Sonquist, J. A. (1963). Ploblems in the analysis of survey data, and a proposal. Journal of the American statistical association, 28, 415-435.（AIDという言葉は見当たらないが，AIDのアルゴリズムが書かれている．）
- ◆ Sonquist, J. A. & Morgan, J. N. (1964). The detection of interaction effects. Survey research center monograph, 35, Institute for social research. The university of Michigan.（AIDversion2に関する報告．ただし，AID3に関するドキュメント（Sonquist, J. A. & Baker, E. L. & Morgan, J. N. (1971) によると，この文献は，”the original AID monograph”と称されている．）
- ◆ 1975年にはCHAID（Kass, 1975）（注1）が登場．その他さまざまなAIDファミリについては，Hawkins & Kass, 1982（注2），など参照．

注1. Kass, G. V. (1975). Significance testing in, and an extension to automatic interaction detection. Applied statistics, 24, 178-189.（この文献では chi-square-based automatic interaction detection の基礎となる考え方が提示され，χ^2分布で近似できる分布をする統計量により，AIDに推定的要素を取り込むことを提案している．しかし，これがいっぱんにCHAIDの文献（Kass, 1975）とされる文献であるかは不明．）

注2. Hawkins, D. M. & Kass, G. V. (1982). Automatic interaction detection. Topics in applied multivariate analysis. ed. D. M. Hawkins, Cambridge university press, 269-302.（Douglas M. Hawkins 編著，医学統計研究会訳（1988）交互作用の自動検出，多変量解析の理論と実際．MPC，283-323．）

方法

■ 分析方針

- ◆ 本稿では，AID（Automatic Interaction Detector）を行った．
- ◆ 従属変数は，スーパー利用回数という変数を新たに作成して利用した．これは，メインデータ.xlsxの255～261列，F_SP_ION3500～F_SP_OTH3500，の17変数を使った．具体的には，それら17変数の和を算出して，間隔尺度の変数とした．
- ◆ 独立変数は，以下2つのファイルの全変数とした（一括して順序尺度として扱った．01型の変数や，都道府県コードなどの名義変数もあるが，変数のひとつひとつについて尺度を指定するには，独立変数が多すぎるため．なお，SampleIDは使用しなかった）（注3）．
  - Webアクセス頻度（0401-0406）
  - 雑誌閲読状況
- ◆ ファイルが大きいため，計算は2段階で行った（注4）．
  - 第1段階は，2つのファイルそれぞれ別々にAIDを行った．
  - 第2段階は，第1段階で木を形成に使われた独立変数を2つのファイルから抜き出して合併し，あらたに1つのファイルを作成し，あらためてAIDを行った．

注3：2つのファイルそれぞれ，従属変数のみ，共通して，スーパー利用回数（17変数の和）の1変数を追加して使用した．

注4：第1段階の計算ではF>2となる独立変数を選び出した．第2段階の計算ではF>13となる独立変数で木を成長させた．

■ ここで使用したAIDについて

◆ F値で木の成長を制御

- 実行の都度，木を成長させるF値の大きさは指定した．（本稿では F > 2 あるいは F > 13 とした．F値がこれを下回る場合は，その枝の成長は終了．）
- 独立変数の各値を水準とみなして（順序尺度として扱った），従属変数の分散，比を計算した．（従属変数は量的変数（間隔尺度）とみなした．）

◆ 3分割以上があり得る

- 独立変数の値は一律に順序尺度であると想定して並び替えた．その後，となりあう水準間で2群の平均値の差の検定（t検定あるいはWelchの検定）を行い，2水準間で従属変数が有意に違う場合（有意確率は，第1段階では5%未満，第2段階では20%未満とした），そこで分岐とした．これを順にとなりあうすべての2水準間で行い分岐を決めた（3分割上も起こりうる．）
- 水準間に指定確率以上に有意となる分岐がない場合も，指定したF値を上回るあいだは，もっともF値が大きくなるように2分岐は行った．
- もっともF値が大きくなるように2分岐しても，指定のF値以下であった場合，そこで枝の成長を止めた．

◆ おなじ説明変数を2度以上使わない

- 一度，分割に使われた独立変数は，それ以降の分割には使用しないものとした．
  - 繰り返し同じ変数が出てくる木は解釈が複雑になるため．木の形成上は正しくないが，木の形成に有効な変数を選ぶという観点では，一定程度の有効性はあるであろうと考えた．この点は踏まえて結果を解釈する必要がある（木の構造は，あらためてこの制約を外して行うべき）．

結果

■ 変数の絞り込み

- ◆ まず，2つのファイルそれぞれについて，独立変数を，F＞2以上となった変数のみに絞り込んだ.
  - 雑誌閲読状況：662変数→222変数
  - Webアクセス頻度（0401-0406）：3907変数→1611変数
- ◆ 絞り込んだ2つのファイルを合併し，合計1,833変数を独立変数として，あらためてAIDを行った．この際，独立変数はF値が13を越えるものだけとし，分岐はp＞0.2となる水準間だけとした.
- ◆ その結果，以下の10変数が，分岐の変数として上がった．（表1）

表1. 決定木（木の成長：F＞13，分割：p＞0.2）を構成した予測変数.

ノード番号	予測変数名	ラベル
ノード(1)	MZ.2102302	オレンジページ（2月2日（土）発売号）
ノード(2)	MZ.2005403	週刊ポスト（3月11日（月）発売号）
ノード(3)	@1808	カルピス
ノード(4)	@0044	ワコール
ノード(5)	@1315	テンプスタッフ
ノード(6)	@1149	三井ダイレクト
ノード(7)	@0677	Yahoo!
ノード(8)	MZ.1710104	日経トレンディ（3月4日（月）発売号）
ノード(9)	MZ.3LVN303	リビング新聞（3月16日（土）配布）
ノード(10)	@2977	ワールド

■ 決定木（図1）

- ◆ いずれも，「0」より「1」で値（スーパー利用回数）は大きい．
- ◆ 分岐に使われた10変数のうち，ON LINE であるものは，以下6つ．
  - ■ カルピス（乳酸菌飲料）
  - ■ テンプスタッフ（一般労働者派遣事業）
  - ■ Yahoo!（ポータル）
  - ■ ワコール（女性下着）
  - ■ 三井ダイレクト（個人向け自動車保険）
  - ■ ワールド（アパレル）

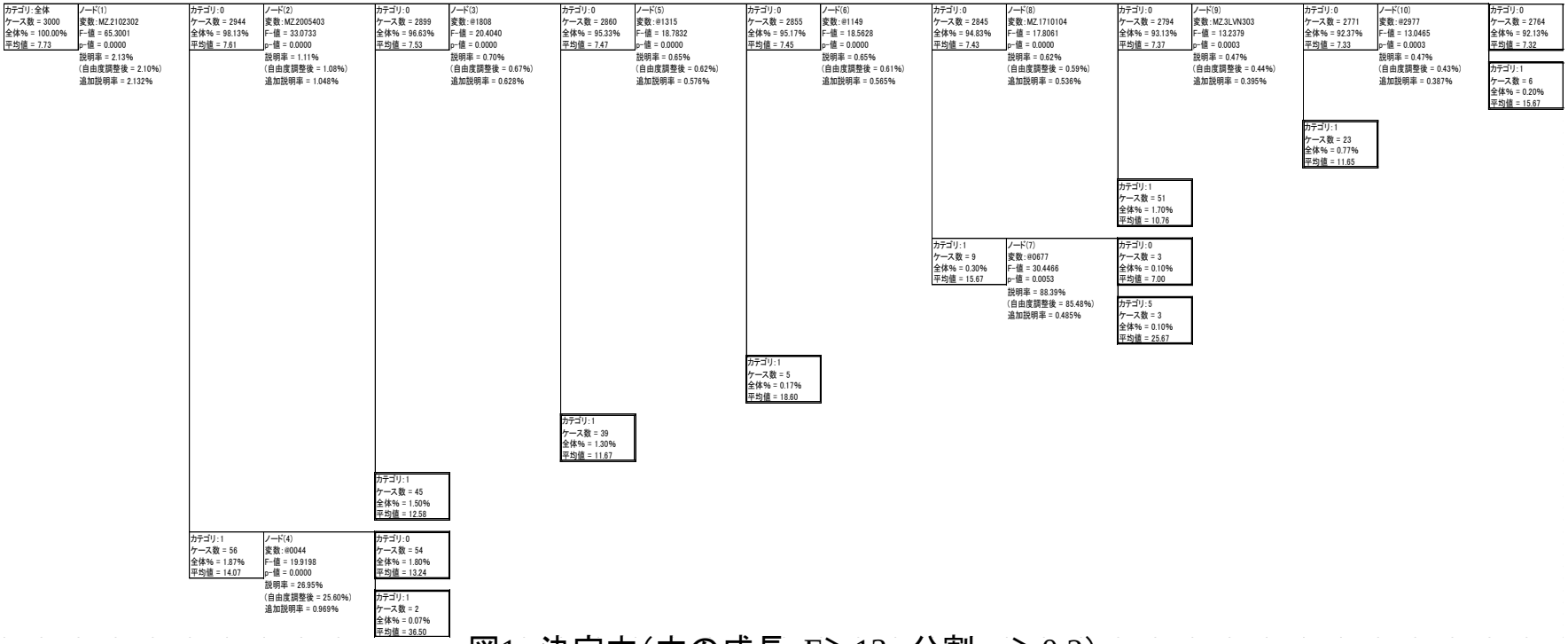

図1. 決定木（木の成長：F＞13，分割：p＞0.2）

考察

- オフラインでの買い物を回数多く行う層が良く訪れるサイト
  - 本稿では，属性を見ていないが，下記のようなオンライン上の行動が見られる層において，よりスーパー利用回数が多い傾向が見られた.
    - カルピス（乳酸菌飲料）のサイトを見る
    - ワコールやワールド（女性下着やアパレル）のサイトを見る.
    - テンプスタッフ（派遣）のサイトを見る.
    - 三井ダイレクト（自動車保険）のサイトを見る.
- webでスーパーの広告を打つ場合は，上記のようなサイトへの出稿も検討の余地があるかもしれない.
  - その場合の期待される効果は，たとえば，以下の通り.
    - カルピスのwebサイトにアクセスする人は，しない人より，7.47回→11.67回（期間単位不明）スーパーを利用する回数が多い.
    - カルピスのwebサイトを訪れる人が行かないwebサイトに広告を出す場合より，カルピスのwebサイトに広告を出すほうが，スーパーを1.56倍使う層に対して広告効果を期待できる.（スーパー利用回数は多いことが期待できるが，利用金額については不明.）
- オフラインの行動にも注目が必要
  - 本稿では注目しなかったが，雑誌購読パタンの違いによっても，webサイト訪問パタンの違いと同程度かそれ以上の効果ができる可能性も示唆された.

■ 検定の多重性

- ◆ 本分析では，いくつか，検定の多重性の問題がある．したがって，本稿で出てくるp値には，あまり確率としての意味は考えるべきではない．
  - ● 分岐の判断には2群の平均値の差の検定を使用したが，これは，まず，2群間の等分散性の検定（両側，$p<0.2$）を行い，その結果を受けてふたたびt検定あるいはWelchの検定を行った．
  - ● 分岐の判断のための平均値の差の検定は，すべての隣り合う2水準間で繰り返して行った．この際，設定した有意確率は，比較あたりの有意水準として用いた．

■ 木の構造が正しいかは不明

- ◆ 本分析で用いたAIDでは，ひとつの独立変数は最大でも1回しか木の成長の過程で登場しない．ひとつの変数が繰り返して木を構成しうる場合，また違った変数が有効な分岐を成す可能性がある．決定木は交互作用の検出を目的の一つとしている点で，この仕様は，交互作用の検出に制限を加え得るもので好ましくない．
- ◆ 本稿では，この点を軽視して，「有効な変数を選び出す」「解釈を複雑にしない」点を重視している点で，有効性に不明瞭な点が残る．ただし，この手法は，実用上，複雑に木が成長して解釈に困ることが少なくない．これは，AID適用自体を敬遠させかねない．この点においては，独立変数を繰り返し分岐に使わないという制約は，わかりやすい，使いやすい，という利点がある．

参考文献

- Hawkins, D. M. & Kass, G. V. (1982). Automatic interaction detection. Topics in applied multivariate analysis. ed. D. M. Hawkins, Cambridge university press, 269-302. (Douglas M. Hawkins 編著, 医学統計研究会訳 (1988)交互作用の自動検出, 多変量解析の理論と実際. MPC, 283-323.)
- Kass, G. V. (1975). Significance testing in, and an extension to automatic interaction detection. Applied statistics, 24, 178-189.
- Morgan, J. N. & Sonquist, J. A. (1963). Ploblems in the analysis of survey data, and a proposal. Journal of the American statistical association, 28, 415-435.
- Sonquist, J. A. & Morgan, J. N. (1964). The detection of interaction effects. Survey research center monograph, 35, Institute for social research. The university of Michigan.