

価値観と行動の相関に注目した消費者の潜在クラス分析

A Latent Class Analysis using Association between Consumer Values and Behaviors.

西尾 義英^{*1}

Yoshihide NISHIO

^{*1} シナジーマーケティング株式会社
Synergy Marketing, Inc.

Consumer segmentation is important in marketing. Consumers can be segmented by their psychological attributes or clustered with their purchase behaviors. But can we make segmentation both psychological and behavioral? We propose a latent class analysis using association rules between consumer values and behaviors. We show such latent classes balances consumer characteristics on values and behaviors in 4 test cases.

1. はじめに

マーケティングにおける消費者のセグメンテーション手法には古典的な年代や性別といったデモグラフィック属性によるものから心理的属性や行動履歴に基づく高度なものまで存在する。中でも我々は消費者の価値観とその類型化に注目して、Societas[谷田 2013]という分析フレームを開発し、クライアント企業に活用いただいている。Societas では 2018 年 3 月現在、12 種の類型と 23 種の価値観変数¹および年代・性別・未婚婚・子どもの有無という 4 種のデモグラフィック属性を規定している。これらはさまざまなクライアント企業が持つデータを匿名状態でつなぐための共通変数の役目を果たす。これにより我々のクライアントは顧客について自社だけでは得がたい情報を得ることができる。

Societas の活用とは、まず 12 種の類型から一つまたは複数の類型をターゲットと定め、その特性を知ることによってアプローチを変えるということが基本である。ただし 12 種では多いあるいはより細かくしたいという場合など、クライアント独自の類型を定義する場合もある。このとき注意すべきことが 2 点ある。まず、クライアントの視点から類型同士の違いが明らかであるということが欠かせない。これは類型の違いでアプローチを変えたいという理由による。クライアントの視点からということは、クライアントが独自に持つデータ、例えば商材の購買傾向において差が大きいと言い換えることができる。もうひとつは我々にとって重要な事柄であるが、類型がクライアントに閉じてしまわないようにしたい。クライアント固有のデータを良く分類できるように最適化しすぎると、共通変数たる Societas との紐付けが難しくなり、他で活用しにくくなる懸念されるためである。整理すると、類型を定義したいデータセットに共通変数とそれ以外の変数が存在する時、それ以外の変数を良く分類しつつ、共通変数から類型を予測できる余地を残したいということである。

従来は価値観の側を因子分析で、あるいは消費行動を潜在クラス分析で、のようにどちらか一方だけから類型を決めることが多かったが、経験的には類型に寄与しない変数との相関はそこまで高くない。また過去にはアンケートから導出されたライフスタイル因子と、ID-POS データに対する商品クラスを階層的に接続したモデルが提案されている[石垣 2011]が、シンプルに価値観と消費行動の両方をバランス良く説明する類型が求められないだろうか。

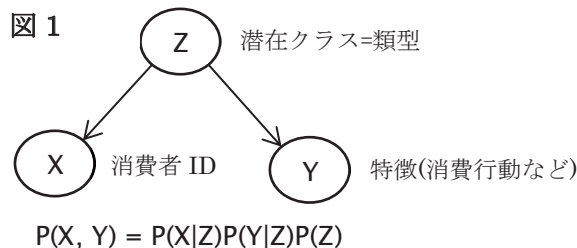
本稿では、価値観と相関する消費行動の組を素性とした潜在クラス分析(pairs-plsa)が我々の求める類型の性質を満たすかを検証する。価値観のみに対する潜在クラス分析(values-plsa)と消費行動のみに対する潜在クラス分析(facts-plsa)を比較対象とし、pairs-plsa が消費行動の説明力において values-plsa より優位で、価値観による類型の予測精度において facts-plsa を上回することを示す。

2. 手法

本稿で行う潜在クラス分析について説明する。

2.1 確率的潜在クラス分析(pLSA)

本稿では図 1 のように類型を潜在クラスと見なし EM アルゴリズムでパラメータ $P(X|Z)$, $P(Y|Z)$, $P(Z)$ を最適化する pLSA[Hofmann 1999]を用いる。ある消費者 X が特徴 Y を持つ確率 $P(X, Y)$ は類型 Z にのみ依存すると仮定するが、これは消費者の特徴を類型によって説明するという用途から自然な仮定といえる。なお本来は複数の潜在クラスの混合によって X と Y の関係を定式化する。これはいわゆるソフトクラスタリングに相当するが、実用上は消費者が唯一の類型に対応すると考えるのが扱いやすいため、 $P(Z|X)$ が最大となる潜在クラスを一つ当てはめることとする。



2.2 クラス数の決定方法

pLSA ではクラス数 K を分析者が与えることとなっているので、同じ条件でクラス数を決定できるように基準を定めておく。AIC や BIC などの情報量基準を用いる例が多い。本稿では AIC を採用する。また EM アルゴリズムの初期値依存性を考慮し、初期値をランダムに変えつつ K を探索する。具体的には異なる K 毎に 10 回繰り返した時の最良解を採用することとする。なおデータセットによっては K を増やしても AIC が悪化する一方ということがありえる。実際今回の実験では全てのデータセットにおいて

連絡先: 西尾 義英, シナジーマーケティング(株),
nishio.yoshihide@synergy101.jp

て提案手法である pairs-plsa のみが AIC で K を選択することができたので、他の手法に対しては同じ K を採用することとする。

2.3 データ形式の説明

本稿での検証では同一の X すなわち消費者 ID の集合に対し Y を 3 パターン作成して pLSA を実行した結果を比較する。以下に各パターンのデータ形式を説明する。

(1) 価値観(values)

Societas が規定する 23 種の価値観変数を用いる。各変数は「好奇心が強い」といった属性を持っているか／持っていないかの二値の確率変数であるが、 $p>0.5$ である属性を列挙したものを Y とする。

(2) 消費行動(facts)

購買データであれば購入した商品の名称やカテゴリを列挙したもの、アンケートの場合は複数選択式の設問で、選択された項目を Y とする。

(3) 価値観と行動の組(pairs)

アソシエーションルール：
rule1 = {価値観属性=ある ⇒ 消費行動=ある} または
rule0 = {価値観属性=ない ⇒ 消費行動=ある}
を Y とするが、データ量が多くなるので関連のある組み合わせのみに絞込む。facts の出現率が比較的高いデータセットにおいては、values と facts の相関係数の絶対値 >0.1 を条件とする。facts の出現率が低いデータセットでは、rule1 と rule0 のリフトを計算し、 $lift(rule1) > 1.1$ かつ $lift(rule0) < 0.9$ となるものだけを残す。

3. 実験

3.1 用いたデータ

実験には性質の異なる以下 4 種のデータセットを用いた。(1) と(2)はインターネット調査によるものである。N は人数、P は変数の数、K は潜在クラス数を表す。

表 1

#	description	N	P	K	pairs の抽出条件
(1)	食についての習慣や意識を問う調査	994	84	14	相関係数
(2)	地方紙とフリーペーパーの購読状況調査	3056	44	10	相関係数
(3)	観光スポットのスタンブ履歴	733	97	5	相関係数 ※しきい値を 0.03 に変更
(4)	スーパーの ID-POS データ	7651	2256	18	リフト

3.2 類型による消費行動の説明力を比較

ここで説明力とは 2.1 で説明した潜在クラス分析をクラスタリング手法と見たときの性能という意味で用いる。まず異なる類型の間で消費行動の特性が異なるということと、同じ類型の中では揃っているという観点で Calinski-Harabasz 指標[式(1)]による比較を行う。

$$CH = \frac{N-K}{K-1} \times \frac{BGSS}{WGSS} \quad \text{式(1)}$$

式 1 の分子 BGSS: Between Group Sum of Square は類型間の分散を表し、分母の WGSS: Within Group Sum of Square は類型内の分散であるため、CH は大きいほど良い。結果を表 2 に示す。

表 2

dataset	Y の形式	BGSS	WGSS	CH
(1)	values	629	13,544	3.50
	pairs	1,477	12,696	8.77
	facts	124	14,049	0.97
(2)	values	171	20,271	2.86
	pairs	2,814	17,627	54.03
	facts	3,177	17,265	62.27
(3)	values	11	2,034	0.97
	pairs	212	1,833	21.04
	facts	235	1,810	23.61
(4)	values	5,624	1,307,416	1.93
	pairs	24,236	1,288,804	8.44
	facts	3,922	1,309,118	1.76

3.3 価値観による類型の予測精度を比較

潜在クラス分析の結果を正解ラベルとし、価値観だけの情報でどれほどの予測精度が得られるかを確認する。ランダムフォレストを分類器に用い、学習データとテストデータを半数に分けた時のラベルの一致率を表 3 に示す。

表 3

dataset	Y の形式		
	facts	pairs	values
(1)	0.301	0.600	0.476
(2)	0.303	0.450	0.610
(3)	0.294	0.338	0.730
(4)	0.251	0.311	0.596

4. まとめ

二つの実験 3.2 と 3.3 からはいずれも仮説を裏付ける結果が得られた。結果について考察し、今後の課題を述べる。

4.1 考察

表 2 では WGSS の差は小さいが BGSS が values<pairs なため CH において values<pairs となっている。消費行動において類型の特徴がはっきり出ていると解釈できる。(1)と(4)では BGSS と CH が facts<pairs となっているのは意外な結果であるが、K が最適ではなかったからかも知れない。

表 3 からは全てのデータセットで facts<pairs、つまり価値観による類型の予測精度が消費行動のみを対象とした潜在クラス分析を上回ることを確認できた。ただし(3)と(4)の予測精度はあまり高くない。

2.2 でも述べたが、今回取り上げたデータセットでは消費行動のみから潜在クラス分析を行う際に AIC 基準で K を選択できなかった。そのようなデータセットであっても価値観と組み合わせ

ることで K が決められたということは提案手法の副次的な利点と言える。

4.2 今後の課題

2.3(3)で述べた組み合わせの数を絞り込む手順において、相関係数やリフトに対するしきい値の選び方には明確な根拠が無いので、統計的な有意性あるいは類型の定量評価を最大化できるような基準を検討すべきであろう。またこの手法から得られる類型の有用性については今後実務における評価を重ねていきたい。

参考文献

- [谷田 2013] 谷田 泰郎, 馬場 彩子, 西尾 義英: *Societas* と社会知ネットワーク, ヒューマンインタフェースシンポジウム 2013, 2013
- [齋藤 2016] 齋藤 有紀子, 木虎 直樹, 谷田 泰郎: 複数データのマッピングによるシニア価値観分析の試み, 2016 年度人工知能学会全国大会 (第 30 回), 2016
- [石垣 2011] 石垣司, 竹中毅, 本村陽一: 日常購買行動に関する大規模データの融合による顧客行動予測システム: 実サービス支援のためのカテゴリマイニング技術, 人工知能学会論文誌, Vol.26, No.6, pp.670-681, 2011.
- [Hofmann 1999] Hofmann, T.: Probabilistic latent semantic analysis, Proc. of Uncertainty in Artificial Intelligence, pp.289-296, 1999.

¹ 23 種の価値観変数: *Societas* の発表時から価値観変数の見直しを行った。現在の変数定義は[齋藤 2016]表 2 を参照のこと。