

はじめに

クラスター分析は便利でポピュラーな分析手法として企業のマーケティング調査データ分析に頻用されている。しかしどの手法も万能ではないように、分類問題に関しても、実際の利用場面で課題を抱えることがある。ユーザーの立場で何点か指摘してみる。理論的には解決されているのかも知れないし、誤用があるかも知れない。

## 1. 消費者のセグメンテーション

マーケティング調査における典型的な分類問題は消費者セグメンテーションである。日本人の半数以上に普及しているような商品、またそのような大きな業界では、母集団を日本全体と設定して、時系列で大規模な消費者調査を実施している。購入商品などの基礎項目のほか、商品開発やコミュニケーション戦略に役立つように消費選好・生活形態・価値観などの質問から消費者のクラスターを作成する。そのクラスターはデモグラフィック属性との関連分析によって記述され命名され、マーケティング上の戦略ターゲットとして運用される。ここでは非階層的手法を使用している。

時系列調査では翌年の調査データは今年のシードを使って割当される。そしてクラスター規模の変化や、他の変数とのクロス分析の結果の変化が解釈されて、アクション計画につなげる。しかし初年度のクラスターは不変でよいのだろうか。翌年はクラスター自体が変化している可能性があるし、そもそも初年度のクラスターが数年間にわたり安定しているだろうか、という疑念を分析者は持つ。一方、マーケティング部門では毎年、クラスターが変わるようでは運用するうえで方針が定まらないので、困る。

クラスターの個数を確定する参照指標はなにがよいのか。あるいは目的別にどの指標がなぜ適切なのかというガイドラインはないだろうか。クラスターの個数の検討と同時に、異なる初期値からの異なる結果の比較・確定の指標も同様である。

クラスターの個数を検討する過程で、サイズが「無意味に」小さいクラスターが出現する。そこで、やや多数のクラスターを作成したうえで、その重心を使って階層的クラスタリングをする。最終的にまとめたクラスターには極端に小さいクラスターはないが、近隣クラスターに含まれている。このようにして再計算された重心が経年的に使用されるが問題はないだろうか。外れ値の影響はどのように、どの程度に深刻なのか頑健なのか。昨年の外れ値の状況と、今年の外れ値、来年の外れ値の状況は同じか違うか。しかも多次元空

間の外れ値の認識は容易ではないから、検討そのものが難しい。

## 2. エリア・マーケティングとエリアの分類問題

個人情報を使わずに目標顧客に接触するために（DM や営業）、ジオデモグラフィック・コードを利用することがある。これらは国勢調査データを分析して、小地域を要素として日本全国で統一したクラスターを作成したものである。代表的なコードとして CAMEO, mosaic, Chomonix などがある。各クラスター・コードの特徴が記述されていて、マーケティング戦略の目標とする小地域を選ぶことができる。たとえば富裕層の多い地域に高額商品を勧めるという戦略を採用することができる。

### CAMEOグループコード

グループ名	各グループの特徴
1. 裕福な単身・2人世帯の多い都会地域	都市部のディンクス中心。上昇志向が強い。
2. 裕福な中高年の多い地域	資産が豊富で老後を楽しむ人々が多い。
3. 裕福なファミリーの多い地域	都市部のファミリーで、子供の教育に熱心。
4. 比較的裕福な単身者の多い地域	将来グループ1や3になる可能性の高い予備軍
5. ホワイトカラー・2世帯住宅の多い地域	ホワイトカラーの多い地域。
6. 平均的な若い単身者の多い地域	若い世代を中心としたグループ
7. 平均的な中高年の多い地域	中高年を中心としたグループ。価格に敏感な傾向。
8. 地方・郊外の単身・2人世帯の多い地域	地方の若い世帯が多い。
9. 地方の中高年・高齢者の多い地域	素朴な生活をたのしむ地方・過疎地域の世帯
10. 農村部など都心から離れた地域	農林業が主要な地方の世帯

### CAMEO詳細コード

グループ名	NO.	詳細グループ名
1. 裕福な単身・2人世帯の多い都会地域	11	都市部の高学歴、単身2人世帯が多い地域
	12	都市部のマンション居住の裕福な専門職世帯が多い地域
2. 裕福な中高年の多い地域	21	自家所有する大きな家に住む裕福な高齢者世帯の地域
	22	地方・郊外の自家所有の戸建に住む裕福な高齢者世帯の地域
3. 裕福なファミリーの多い地域	31	都心部の小規模マンションに住む世帯が多い地域
	32	就学児童をもつファミリーが多い地域
	33	マンションに居住する専門職世帯が多い地域
	34	自己所有の戸建とマンションに居住する世帯が混在する地域

たいへん便利であるが、このように分類された結果を利用すると、以下のような問題に直面する。

- (1) 関東とか関西というように地域を絞ってみたら、同じようなクラスターばかりになることがある
- (2) それと関連するが、その地域では、あまりにも小さなクラスター・サイズになってしまって使い物にならない場合がある
- (3) 同じクラスター・コードなのに、関西と関東ごとに内容を分析したら、同じ消費者像とは思えない地域がある

これは全国ベースでは特徴の変動がクラスターに寄与しているが、データの地域範囲を限定することで特徴が失われる（集中する）ためである。解決策は、関東データに絞ったうえで再分析することである。すると今度は東京という範囲に絞ったら同じ現象が生じる。データを絞る範囲でクラスターの結果は変わってしまいキリがない。採用する変数にも大きく依存する。結局、クラスターには一貫性がなくて、いつでも「新鮮な」分析をすることになる。

クラスターは平均的傾向を示しているが、個別の小地域を個別に確認していくと、どうしてもそのクラスターらしくない地域も含まれている。重心からの距離が遠い要素だと想像される。解決策としてはクラスター重心で再度、階層的クラスタリングをして近隣への再分類を検討することだが、「分類した、そのあと」の分類のためにはクラスター間距離や分析変数の内容が正確に公表されていなければ対処できない。

統計法の改正で国勢調査のオーダーメイド集計ができるようになった。商用には許可されないが学術目的には利用できる。すでに集計された表側と表頭の質問間クロス集計がサービスされているが、小地域の分析では生データの単位から検討が必要である。

### 3. テキストマイニングとテキストの分類問題

クチコミ分析では言葉の分類問題がある。ブログ解析などで、ある商品についての言及を調べるのだが、ポジティブな内容か、ネガティブかを分類（判別）したい。また「何について、どのように語られているのか」を分類して集計したい。

どのようなソフトウェアにもあるのは単語の出現頻度ランキングである。いわば次元の分析だが、これで分ることは少ない。

もう少し発展した分析になると「係り受け」分析があるが、2対にすると件数が非常に減ってしまって全体が描けず、バラバラになってしまう。たとえば単語だけなら「ラー油」3824件であるが、「ラー油・食べる」の係り受けは485件だけになってしまう。

No.	単語	品詞	件数
1	桃ラー	名詞	11,935
2	ラー油	名詞	3,824
3	桃屋	名詞	3,380
4	食べる	動詞	2,549
5	RT	名詞	2,177
6	買う	動詞	1,850

No.	単語	品詞	件数
1	ラー油 - 食べる	名 - 動	485
2	桃ラー - 食べる	名 - 動	440
3	桃ラー - 買う	名 - 動	397
4	ラー油 - 買う	名 - 動	193
5	桃ラー - 入れる	名 - 動	171
6	桃ラー - 美味しい	名 - 形	157
7	桃ラー - ゲットする	名 - 動	156
8	桃ラー - 売る	名 - 動	156

単語の分類は、一方の極では人工知能的に「自動」判定するアルゴリズムであり、こちらは誤分類の少なさを競っているわけが、誤分類を減らすと見逃しが増えてしまうことから、見逃しを無くそうとすると必ず一定比率の誤判定が混ざってくる。もう一方の極は「手動」作成である。この作業は分析にいたる以前に準備作業だけで体力が尽きてしまうほど大変である。実際には自動判定と手動作成の中間的な支援（サポート）機能が欲しい。

「高い」「低い」という分類をしたら、次にそれらが「価格」についてか「画素数」についてなのかを高次に分類するような段階的な分類作業を確認しながら、ある程度は自動的に分類する機能が期待される。いわば2次元に視覚的に分類を配置するような機能である。

#### 4. データマイニングと安定性・再現性・効率性の問題

WEB上で蓄積される購買履歴データは毎日1万件、年間360万件、3年間分析すると1000万件になる。購買者と購入商品の多変量データから消費者を分類したい。特殊な高性能コンピュータではなく、企業内で使用されるハイエンドなパソコンとマイニング・ソフトウェアを前提とすると、実際には大規模データを分析するのは難しい。

ハードウェアの性能問題としては分析を繰り返し検討・吟味する観点からは1～2時間であれば実用的だが、10～30時間もかかってようやく結果が出る状況であり、これでは業務に支障をきたし事実上、使えないということになる。K-meansのような手法でこのような状況であり、潜在クラス分析やMCMCは実際には使うことができない。ソフトウェア、アルゴリズムの問題として、すでに解決されているのだろうか。

機動的に分析結果を反復しながら検討できないうえに、分類結果が一意に定まらない問題がある。観測値が多だけでなく、変数も多い。自然なクラスターが存在していてそれを発見するという状況ではなく、明確な境界のない消費者購買データ空間を適切に分割するという状況である。異なる初期値に対して、解釈可能な異なる分類結果がもたらされる。消費者分類なので分類結果に唯一の正解がないことが自然な状況であるともいえる。このような状況に分類問題はソリューションを持っているだろうか。

データマイニングの変数が多い。重回帰分析や因子分析において変数選択が議論され、選択法が存在しているように、クラスター分析における変数選択問題は解決されているのだろうか。分析者の主観や業務知識や洞察力で選択されているのが現状ではないだろうか。