単相3元重複クラスター分析法とその応用に関する研究

○横山 暁*　　中山 厚穂**　　岡太 彬訓***

*慶應義塾大学大学院理工学研究科　**立教大学経営学部　***多摩大学大学院経営情報学研究科

1　はじめに

近年の情報処理技術の発展により，様々な分野で様々な種類のデータ，特にいくつもの対象や変量，属性間の関係を表す多元データが容易に取得できるようになった．しかし，親近度データの分析では，単相 2 元データに個人差を含めた 2 相 3 元データの分析以外，3 元データや多元データの分析はあまり研究がされておらず，クラスター分析法においては，3 相 3 元データに適用できるモデルが提案されている程度であり，単相 3 元データの分析はほとんど分析されていない．

本発表では，Yokoyama et al. (in press) で提案された単相 3 元重複クラスター分析法を紹介するとともに実際の分析例についても紹介した．

2　単相 3 元重複クラスター分析法

2.1　既存研究

重複クラスター分析法は Shepard and Arabie (1979) によって単相 2 元データに適用できるモデル (ADditive CLUStering, ADCLUS) が提案され，そのアルゴリズムを改良した Arabie and Carroll (1980) による MAthematical Programming CLUStering (MAPCLUS)，および 2 相 3 元データに適用できるように拡張した Carroll and Arabie (1983) の INdividual Differences CLUStering (INDCLUS) が代表的な手法である．

ADCLUS のモデルでは 2 つの対象間の類似度 s_{ij} （ただし $i, j = 1, \ldots, n$，n は対象の個数）を，対象のクラスターへの所属を表すパラメータと，各クラスターの重みを用いて

$$s_{ij} \cong \sum_{r=1}^{R} w_r p_{ir} p_{jr} + c \tag{2.1}$$

で表す．ここで，$w_r (r = 1, \ldots, R)$ はクラスター r に対する非負の重み（係数）であり，p_{ir} は対象 i がクラスター r に所属すれば 1，そうでなければ 0 となる．また，c は加算定数である．

MAPCLUS のアルゴリズムは ADCLUS のモデルに適用するものであり，ADCLUS で提案されたアルゴリズムに比べて計算手続が大幅に改良されている．この MAPCLUS のアルゴリズムでは $\boldsymbol{P} = [p_{ir}]$ を推定するために損失関数を定義し，VAF 比が最大になるように勾配法を用いて損失関数を最小化する方法であり，離散問題を解くことを，制約付きの連続問題とみなす数理計画法で構成されている．

2.2　単相 3 元重複クラスター分析法

Yokoyama et al. (in press) で提案された単相 3 元重複クラスター分析法のモデルでは，単相 3 元類似度データ $\boldsymbol{S}^{(3)} = [s_{ijk}]$ （ただし $i, j, k = 1, \ldots, n$，n は対象の個数，$\boldsymbol{S}^{(3)}$ は $n \times n \times n$ の 3 元行列）を ADCLUS のモデルと同様に対象のクラスターへの所属を 2 値で表す p_{ir} と，各クラスターごとの重み w_r を用いて

$$s_{ijk} \cong \hat{s}_{ijk} = \sum_{r=1}^{R} w_r p_{ir} p_{jr} p_{kr} + c \tag{2.2}$$

で推定する．ここで $\hat{s}_{ijk}$ は推定された類似度である．本手法のアルゴリズムは，MAPCLUS のアルゴリズムを単相 3 元データに適用できるように改良したものである．

3　データへの適用

Yokoyama et al. (in press) では，あるコンビニエンスストアにおける同時購買データの分析を行っている．この分析では，15 の商品カテゴリー間における 3 カテゴリー以上の同時購買データを抽出し，3 カテゴリーの同時購買の頻度を単相 3 元類似度データとして分析を行っている．さらに，2 カテゴリー以上の同時購買データから単相 2 元類似度データを作成し，単相 2 元重複クラスター分析法との分析結果の比較も行っている．

横山・岡太 (2008) では Web Page のアクセスログデータの分析を行っている．本研究では，横山・岡太 (2008) を基に単相 3 元重複クラスター分析法および単相 2 元重複クラスター分析法での分析結果の違いについて比較・検討を行った．分析に用いたデータは，1999 年 9 月 28 日の 24 時間における msnbc.com へのアクセスログデータである．このデータは 989,818 人の訪問者の 17 のページカテゴリーの閲覧遷移データであり，ページカテゴリーは front page, news, tech, local, opinion, on-air, misc, weather, health, living, business, sports, summary, bbs, travel, msn-news, msn-sports に分類されている．このデータから連続して閲覧されたページカテゴリーに類似性があるとみなし，単相 2 元類似度データと単相 3 元類似度データを作成し分析を行った．分析の結果，どちらの分析でも front page がすべてのクラスターに所属する一方，front page 以外のクラスターの構成要素は異なる結果となり，単相 3 元データを分析することで，単相 2 元データの分析では得ることのできない結果を得ることができた．

4　まとめと今後の課題

本発表では重複クラスター分析法，特に Yokoyama et al. (in press) による単相 3 元重複クラスター分析法について紹介するとともに，実際のデータの分析例についても紹介した．

本手法は従来分析することのできなかった単相 3 元親近度データに適用することができ，従来得られることのできなかった新たな結果の解釈が得られることが判った．今後様々なデータに適用することで，本手法の有効性を検討するとともに，単相 2 元データ分析と単相 3 元データ分析のどちらの結果を採用すべきかという点についても検討を行いたい．

参考文献

Arabie, P. and Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 45*, 211-235.

Carroll, J. D. and Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika, 48*, 157-169.

Shepard, R. N. and Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*, 87-123.

横山 暁・岡太彬訓 (2008). 単相データ分析におけるデータ形式の違いによる分析結果の影響について [要旨]. 日本行動計量学会第 *36* 回大会発表論文抄録集, pp. 161-162.

Yokoyama, S., Nakayama, A., and Okada, A. (in press). One-mode three-way overlapping cluster analysis. *Computational Statistics*.