欠測したアンケートデータに対する局所的数量化分析

呉　　志賢

1. はじめに

アンケートデータのような多変量カテゴリカルデータの分析は、業務や分野を問わず様々な場面で用いられており、個体や項目に数量的得点を与え、それを低次元配置することによって、内在する構造を把握することができる。日本では林による数量化分析3類[1]が有名であるが、等質性分析[2]、西里の双対尺度法[3]なども、本質的には同等の解に至る点で同一の解析法と見なすことができる。これらの手法では、個体や項目の間の関係をよく表す低次元の散布図を作成するために質的データに数量的得点を与えることが目的となるが、個体や項目の数が多い場合には、人間が直感的に意味を捉えられる有用な低次元表現を得ることは難しい。

そこで、不要な個体や項目を取り除いたり、いくつかのグループに分割したりすることにより、低次元配置が可能なグループを見出し、尺度構成を行う方法がいくつか研究されている。土屋[4]は項目を分割することによって、本多ら[5]は個体を分割することによって、局所的な数量化分析法を提案している。また、呉ら[6]は個体と項目の両方を分割する手法を提案している。これらの手法は、大規模なデータベースから潜在的な特徴を見出されることが示されている。ただし、それらのデータベースにおけるデータは欠測値を含むことが多い。本研究では、無回答を含むアンケートからの多変量カテゴリカルデータが与えられたときに、欠測値を考慮しながら、個体と項目の両方を分割して低次元散布図を得る手法を提案する。提案法は、等質性分析[2]に個体と項目の両方に対応したメンバシップを導入することでファジィクラスタリングを施す。また、欠測値に対応する逸脱度に重み0を掛け合わすことにより、すべての観

※　本研究は2005年度大阪経済法科大学研究補助金の成果の一部である。

測値を生かしたクラスタリングを行う。

2．多変量カテゴリカルデータの数量化

本研究では、アンケートにおいて1つの項目に対して数個の選択肢が用意されていて、個体はその中から答えを1つ選択するという形式で収集された多肢選択の多変量カテゴリカルデータを扱う。多変量カテゴリカルデータにおいて、個体数を n、項目数を m として各個体を $i\,(i=1,\cdots,n)$、各項目を $j\,(j=1,\cdots,m)$ と表し、各項目はそれぞれ K_j 個のカテゴリ $k\,(k=1,\cdots,K_j)$ を持つものとする。ある項目 j に対して、ダミー変数 g_{ijk} 及び $(n\times K_j)$ 行列 $\boldsymbol{G}_j$ を以下のように定める。

$$g_{ijk}=\begin{cases}1\ ;\text{ある項目}j\text{について個体}i\text{がカテゴリー}k\text{を選んだとき}\\ 0\ ;\text{それ以外}\end{cases}\tag{1}$$

$$\boldsymbol{G}_j=\begin{pmatrix} g_{1j1} & \cdots & g_{1jk} & \cdots & g_{1jK_j}\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ g_{ij1} & \cdots & g_{ijk} & \cdots & g_{ijK_j}\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ g_{nj1} & \cdots & g_{njk} & \cdots & g_{njK_j}\end{pmatrix}\tag{2}$$

個体と項目を数量化するために、割り当てる数量的得点を定義する。得点は $P\,(p=1,\cdots,P)$ 次元とし、個体および項目に割り当てる得点をそれぞれ $(n\times P)$ 行列 $\boldsymbol{X}$、$(K_j\times P)$ 行列 $\boldsymbol{Y}_j$ にまとめ、次のように表す。個体と項目の数量的得点に基づき低次元散布図を得る。

$$\boldsymbol{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1P}\\ x_{21} & x_{22} & \cdots & x_{2P}\\ \vdots & \vdots & \ddots & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{nP}\end{pmatrix}\tag{3}$$

$$\boldsymbol{Y}_j = \begin{pmatrix} y_{j11} & y_{j12} & \cdots & y_{j1P} \\ y_{j21} & y_{j22} & \cdots & y_{j2P} \\ \vdots & \vdots & \ddots & \vdots \\ y_{jK_j1} & y_{jK_j2} & \cdots & y_{jK_jP} \end{pmatrix} \tag{4}$$

等質性分析[２]は、各個体の得点とその個体が反応した選択肢の得点が等質的な値をとるという原理に基づいて数量的得点を推定する。項目jに関する個体iの等質性の仮定からの逸脱度d_{ij}^2を

$$d_{ij}^2 = \left\| x_{ip} - \sum_{k=1}^{K_j} g_{ijk} y_{jkp} \right\|^2 \tag{5}$$

と定義する。データ全体の仮定からの逸脱度は、項目に関して平均をとり、個体を合計する。これを非等質性基準とする。

$$\sigma = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij}^2 = \frac{1}{m} \sum_{j=1}^{m} \mathrm{tr}\left\{ (\boldsymbol{X} - \boldsymbol{G}_j \boldsymbol{Y}_j)^{\mathrm{T}} (\boldsymbol{X} - \boldsymbol{G}_j \boldsymbol{Y}_j) \right\} \tag{6}$$

３．欠測したアンケートデータに対する局所的数量化分析

個体と項目をC個のクラスターに分割するために、クラスターに対するそれぞれのメンバシップを定義する。個体iのクラスターc（$c = 1, \cdots, C$）におけるメンバシップをu_{ci}とし、項目jのメンバシップをw_{cj}とする。また、u_{ci}を対角要素とする（$n \times n$）対角行列を$\boldsymbol{U}_c = \mathrm{diag}\,(u_{c1} \cdots u_{cn})$とする。

等質性分析は主成分分析における最小２乗基準と類似性があり、主成分分析の質的データへの拡張と位置づけることができる。一方、最小２乗基準を用いた主成分分析の目的関数にメンバシップを導入したものは、Fuzzy c-Varieties（FCV）法[７]の目的関数に帰着できる。それらの類似点に着目すると、ファジィクラスタリングと等質性分析の同時分析法の目的関数を以下のように定義することができる。

$$\tilde{\sigma}=\frac{1}{m}\sum_{c=1}^{C}\sum_{j=1}^{m}w_{cj}\,\mathrm{tr}\left\{(\boldsymbol{X}_c-\boldsymbol{G}_j\boldsymbol{Y}_{cj})^{\mathrm{T}}\boldsymbol{U}_c\boldsymbol{M}_j(\boldsymbol{X}_c-\boldsymbol{G}_j\boldsymbol{Y}_{cj})\right\}+\lambda_u\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci}$$
$$+\lambda_w\sum_{c=1}^{C}\sum_{j=1}^{m}w_{cj}\log w_{cj} \tag{7}$$

ここで、$\boldsymbol{M}_j=\mathrm{diag}\,(m_{1j}\cdots m_{nj})$ は $(n\times n)$ 対角行列で、個体 i の項目 j に関してデータが観測されていれば 1、欠測していれば 0 をもつ 2 値変量 $m_{ij}=\sum_{k=1}^{K_j}g_{ijk}$ を対角要素とする。すなわち、カテゴリカルデータにおいて無反応な項目があった場合は欠測値とみなし、重みを 0 として解析を行うことになる。

式（7）の第 2、第 3 項は、メンバシップのファジイ化のためのエントロピー正則化[8]を表す。λ_u と λ_w はそれぞれ、個体と項目のファジイ度を調節するパラメータである。

ここで、一意な解を得るために、各クラスターにおける個体の得点 $\boldsymbol{X}_c$ について、次の正規化条件を導入する。

$$u_c^{\mathrm{T}}\sum_{j=1}^{m}\boldsymbol{M}_j\boldsymbol{X}_c=\boldsymbol{0}^{\mathrm{T}} \tag{8}$$
$$\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{U}_c\sum_{j=1}^{m}\boldsymbol{M}_j\boldsymbol{X}_c=m\Big(\sum_{i=1}^{n}u_{ci}\Big)\boldsymbol{I}_p \tag{9}$$

$\partial\tilde{\sigma}/\partial\boldsymbol{Y}_{cj}=0$ より、$\boldsymbol{Y}_{cj}$ の解は、

$$\boldsymbol{Y}_{cj}=(\boldsymbol{G}_j^{\mathrm{T}}\boldsymbol{U}_c\boldsymbol{G}_j)^{-1}\boldsymbol{G}_j^{\mathrm{T}}\boldsymbol{U}_c^{\mathrm{T}}\boldsymbol{X}_c \tag{10}$$

となる。$\partial\tilde{\sigma}/\partial\boldsymbol{X}_c=0$ より、

$$\boldsymbol{X}_c=\Big(\sum_{j=1}^{m}w_{cj}\boldsymbol{M}_j\Big)^{-1}\sum_{j=1}^{m}w_{cj}\boldsymbol{G}_j\boldsymbol{Y}_{cj} \tag{11}$$

を得る。また、$\partial\tilde{\sigma}/\partial w_{cj}=0$ と確率的制約 $\sum_{j=1}^{m}w_{cj}=1$ より、

$$w_{cj}=\exp(A_{cj}-1)\Big/\sum_{l=1}^{m}\exp(A_{cl}-1) \tag{12}$$

$$A_{cj}=\frac{-1}{\lambda_w m}\sum_{i=1}^{n}\sum_{p=1}^{P}u_{ci}m_{ij}\Big(x_{cip}-\sum_{k=1}^{K_j}g_{ijk}y_{cjkp}\Big)^2 \tag{13}$$

となる。同様に $\partial\tilde{\sigma}/\partial u_{ci}=0$ と確率的制約 $\sum_{i=1}^{C}u_{ci}=1$ より、

$$u_{ci}=\exp(B_{ci}-1)\Big/\sum_{l=1}^{C}\exp(B_{li}-1) \tag{14}$$

$$B_{ci}=\frac{-1}{\lambda_u m}\sum_{j=1}^{m}\sum_{p=1}^{P}w_{cj}m_{ij}\Big(x_{cip}-\sum_{k=1}^{K_j}g_{ijk}y_{cjkp}\Big)^2 \tag{15}$$

となる。

以下にアルゴリズムを記述する。アルゴリズムには、交互最小2乗法の原理に基づく繰り返しアルゴリズムを採用している。

欠測したアンケートデータに対する局所的数量化分析法

Step1 メンバシップ u_{ci} および $\boldsymbol{X}_c$ を乱数により初期化し、u_{ci} の確率的制約と式（8）、（9）を満たすように基準化する。

Step2 式（10）より、$\boldsymbol{Y}_{cj}$ を求める。

Step3 式（11）より、$\boldsymbol{X}_c$ を求め、式（8）、（9）を満たすように基準化する。

Step4 式（12）より、メンバシップ w_{cj} を更新する。

Step5 式（14）よりメンバシップ u_{ci} を更新する。

Step6 反復回数条件 ε を満たせばStep7へ、それ以外はStep2へ戻る。

4．インターネットに関する欠測したアンケートデータの分析

携帯電話やブロードバンド回線の急激な普及により、インターネットが身近なものになった現代社会において、我々の生活は便利で豊かになるものと期待される。現在のインターネットの普及状況やそれに対するイメージ、今後期待するサービスの動向を考察するため、アンケートを行い、提案法を用いてデータ分析を行った。以下に、分析を行ったデータの概要を記す。なお、データは

いくつかの欠測値を含む。

インターネットに関するアンケートデータ

(1) 調査方法

- 調 査 対 象：15歳以上70歳未満の男女
- 実 施 時 期：平成16年7月1日～9月28日
- 調 査 手 法：アンケート
- 有効回答数：162人（男：90人、女：72人）

(2) 調査内容

- インターネット利用状況
- インターネットに対するイメージ
- インターネットに期待するサービス

(3) アンケート項目

- ユーザー属性：年齢、性別、職業、同居人数
- インターネット利用環境：パソコン所有状況、インターネット接続環境、インターネットと電子メールの利用状況
- インターネットの利用用途：目的、検索、電子メールの使用用途
- インターネットに対するイメージ：開放度、真面目さ、現実性、公平性、健全性、利便性、格好、身近さ、必要性、楽しさ
- インターネットへの期待：利用したいサービス、期待するサービス

4.1 パソコンの所有に関する考察

提案法を用いて、回答者のパソコンの所有状況を調べた。その結果の散布図を図1、2に示す。クラスター1より、家族暮らしの会社員や大学生は共用PCを持ち、一人暮らしの大学院生は学校のPCと個人PCを共に持っているとわかる。また、主婦や自営業に関しては、個人PCを持っているといえる。一方、クラスター2より、会社員は会社PC、一人暮らしの大学生は個人PC、

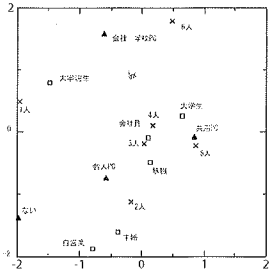

図1 パソコンの所有状況（クラスター1）

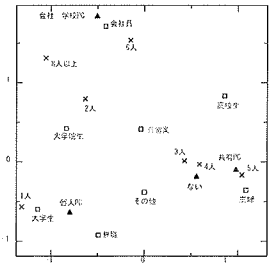

図2 パソコンの所有状況（クラスター2）

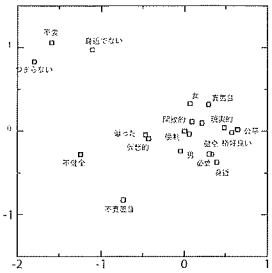

図3 インターネットに対するイメージ（クラスター1）

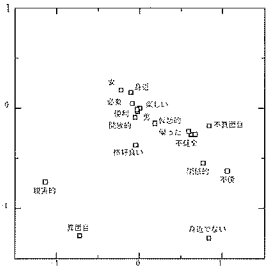

図4 インターネットに対するイメージ（クラスター2）

主婦は共用PCを持っていることが読み取れる。これらより、家族暮らしの大学生を対象にしたクラスター1と一人暮らしの大学生を対象にしたクラスター2に分けることで、パソコンの所有状況をより明確に表せることができた。

4.2 インターネットに対するイメージに関する考察

インターネットに対するイメージは、利用や発展に大きく影響する。インターネットのイメージと性別の関係を調べた。その結果の散布図を図3、4に示す。クラスター1より、男性、女性ともにプラスのイメージを多く持っていることがわかる。クラスター2より、女性は身近、必要といったプラスのイメージに対して男性はそのようなイメージと一緒に仮想的、偏ったといったマイナスのイメージも合わせ持つことがわかる。

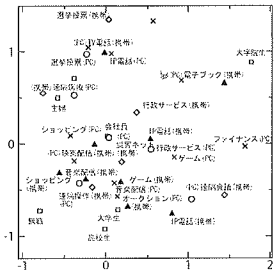

図5　今後期待するサービス（クラスター1）

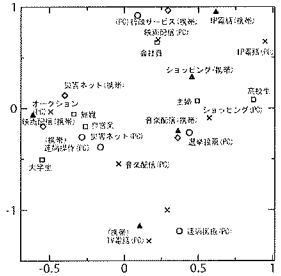

図6　今後期待するサービス（クラスター2）

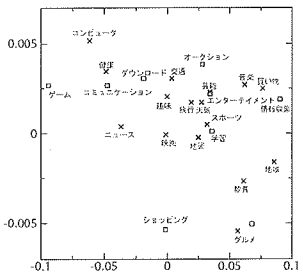

図7　インターネットの検索（クラスター1）

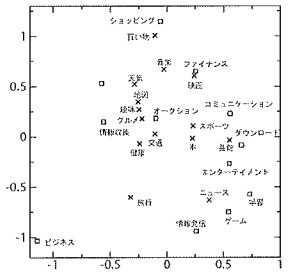

図8　インターネットの検索（クラスター2）

4.3　今後期待するサービスに関する考察

インターネットの今後の展望に関して、利用したい、または期待するサービスと職業との関係を調べた。その結果の散布図を図5、6に示す。クラスター1より、会社員や主婦は利用したいサービスではTV電話やIP電話といった新しいコミュニケーション手段を、期待するサービスでは選挙投票や遠隔医療といった時間や距離に拘束されないサービスを望んでいることがわかる。それに対して、大学生、高校生は利用したいサービスでは映画配信、音楽配信、オークションといった娯楽を、期待するサービスでは遠隔操作といった生活を便利にするサービスを望んでいることがわかる。クラスター2より、会社員は利用したいサービスで映画配信、期待するサービスで選挙投票、主婦は利用したいサービスでショッピング、期待するサービスで行政サービス、大学生は利用し

たいサービスでオークション、期待するサービスで災害ネットを望んでいると読み取れる。これらより、職業ごとに最も望んでいるサービスを、直感的に捉えることができる。

4.4　データ全体に関する考察

最後に、個々の分析テーマごとではなく、新たにデータ集合全体から潜在的な一次元性を抽出することを目的に、全項目に対して提案法を適用した。分析の結果、パソコンを使用してインターネットを行う目的と、よく検索するカテゴリの2つの項目が抽出され、クラスターごとに散布図7、8が得られた。クラスター1ではコミュニケーションを目的とする人はコンピュータ・健康・ニュースを、ダウンロードを目的とする人は趣味・映画、エンターテイメントを目的とする人は芸能・天気・スポーツ、情報収集を目的とする人は買い物・音楽を検索していることがわかる。クラスター2ではショッピングを目的としている人は買い物を、ファイナンスを目的としている人は映画・音楽、情報収集を目的としている人は天気・地図・趣味・グルメ・交通・健康を検索している。以上のことから、提案法によってデータの潜在的な局所的特徴を抽出できることがわかる。

5. おわりに

本論文では、欠測値を含むアンケートデータが与えられたときに、それらの個体と項目の両方を分割しながら知識発見に有用な低次元散布図を得る手法を提案した。提案法により、個体と項目の数が多い場合に、それらを分割することで人間が直感的に意味を捉えられる有用な低次元表現を得ることができると考えられる。多変量カテゴリカルデータの局所的な分析法としては、数量化分析3類にファジィクラスタリングを融合した手法や、個体と項目のメンバシップを凝集度の基準により推定する手法なども提案されており、それらの手法との関連の検証が今後の課題である。

参考文献

[1] C. Hayashi, "On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematical statistical point of view," *Annals of the Institute of Statistical Mathematics*, Vol.3, pp.69-98, 1952.

[2] A. Gifi, *Nonlinear Multivariate Analysis*, John Wiley & Sons, 1990.

[3] 西里静彦，「質的データの数量化—双対尺度法とその応用—」，朝倉書店，1982.

[4] 土屋隆裕，“項目分類のための数量化法”，行動計量学，Vol.22, No.2, pp.95-109, 1995.

[5] K. Honda, Y. Nakamura and H. Ichihashi, "Simultaneous application of fuzzy clustering and quantification with incomplete categorical data," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 8, No. 4, pp.183-193, 2004.

[6] C.-H. Oh, K. Honda, and H. Ichihashi, "Quantification of multivariate categorical data considering clusters of items and individuals," *Modeling Decisions for Autificial Intelligence, Lecture Notes in Artificial Intelligence 3558*, ed. by V. Torra, Y. Nakamura, and Y. Miyamoto, Springer, pp.164-171, 2005.

[7] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum press, 1981.

[8] 宮本定明，馬屋原一孝，向井殿政男，“ファジィ c-平均法とエントロピー正則化法におけるファジィ分類関数”，日本ファジィ学会誌，Vol.10, No.3, pp.156-165, 1998.