HOKKAIDO UNIVERSITY

| | |
|---|---|
| Title | 3-wayデータ数量化の1手法について |
| Author(s) | 山ノ井, 髙洋; 潘, 旅家 |
| Citation | 北海道大學工學部研究報告, 132, 155-160 |
| Issue Date | 1986-07-31 |
| Doc URL | http://hdl.handle.net/2115/42000 |
| Type | bulletin (article) |
| File Information | 132_155-160.pdf |

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

北海道大学工学部研究報告
第132号（昭和61年）

Bulletin of the Faculty of Engineering,
Hokkaido University No. 132 (1986)

# 3－way データ数量化の1手法について

山ノ井高洋*・潘　旅家**

（昭和61年3月31日受理）

# On a quantification method for 3-way data

Takahiro YAMANOI and Lucas PUN

(Received March 31, 1986)

**abstract**

A method, called Hayashi's quantification II, was proposed by Hayashi for 2-way categorical data. The method, proposed in this paper, gives an extension of the quantification II to 3-way categorical data. As well as Hayashi's quantification method II, the method provides a quantification of data for classifying them into patterns.

## 1．緒　　言

カテゴリカル・データ解析の1手法として林の数量化分析がある。[1)2)3)] この手法は主として，個体と1つの観点によるアイテム・カテゴリーからなる，いわゆる2－way データを対象として取り扱っている。なかでも数量化分析II類は，パターン分類に有効で，近年文字認識に応用した例が見られる。[4)5)6)] これらの研究は，従来の物理的特性とは異なる機能的属性に着目し，2－way カテゴリカル・データとして数量化を試みたものである。一方，従来の物理的特性に着目し，文字認識を考え，対象を白黒データに限れば，これらは1－0データが2次元に個体ごとに配列された，一種の3－way カテゴリカル・データとなる。数量化分析III類の拡張としての3－way データ数量化の研究はいくつか見られるが，[7)8)9)] II類の拡張なるものはまだない。本論文では，データ解析の1手法として，数量化分析II類の3－way データへの拡張を試みる。これは，パターン分類の手法として，文字認識への応用も可能である。

## 2．3－way データの定義と数量化の準備

$n$ 個の個体 $P_i(i=1,\cdots,n)$ に関して，2組の観点 $(Q_i,\ R_k)\ (j=1,\ \cdots,\ l:k=1,\cdots,\ m)$ から測定された同時パターン $\delta_i(j,\ k)$ が与えられているものとする。ただし，$\delta_i(j,\ k)$ は

$$\delta_i(j,\ k)=\begin{cases} 1 : P_i \text{ が } Q_j \text{ と } R_k \text{ に同時に反応したとき} \\ 0 : \text{それ以外のとき} \end{cases}$$

と定義する。ここで，$n$ 個の個体は $g$ 個のグループに分かれているものとする。このデータを図示したものが Fig. 1 である。この3－way データに対して，次の分類問題を考える。

問題：与えられた同時パターン $\delta_i(j,\ k)$ とグループの情報をもとに，グループが未知の同時パターン $\delta_h(j,\ k)$ を持つ個体 $P_h$ をグループに分類する。

＊　情報工学専攻　情報数理工学第一講座
＊＊フランス国　ボルドー第1大学　GRAI 研究所

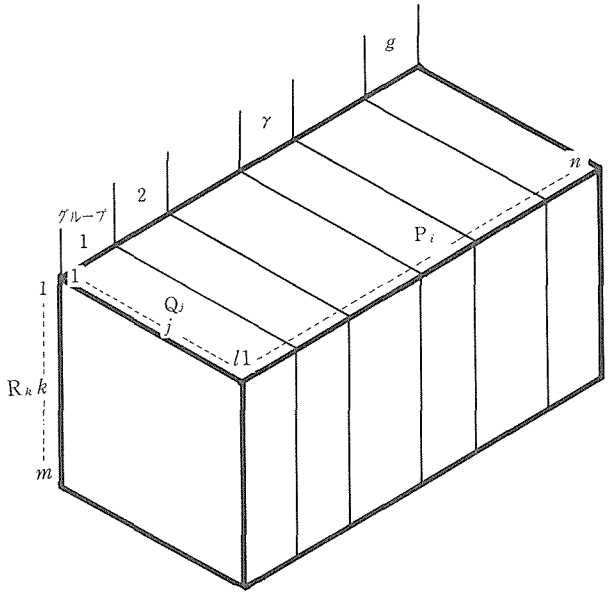


**Fig. 1**　3－way データの概念図

この問題を解くための準備として以下の諸量を定義する。

$$\delta_i(\gamma)\equiv\begin{cases}1: i\text{番目の個体が}\gamma\text{番目のグループに含まれているとき}\\0: \text{それ以外のとき}\end{cases}$$

$$d_i(\beta)\equiv\begin{cases}\sum_{k=1}^{m}\delta_i(j,k): \beta\equiv j\\ \sum_{j=1}^{l}\delta_i(j,k): \beta\equiv l+k\end{cases}$$

$$D_i(\alpha)\equiv\delta_i(j,k),\quad \alpha\equiv m(j-1)+k$$

$$\overline{D}(\alpha)\equiv\frac{1}{n}\sum_{i=1}^{n}D_i(\alpha),\ \delta(\gamma)\equiv\frac{1}{n}\sum_{i=1}^{n}\delta_i(\gamma),\ L\equiv l\times m,$$

$$\bar{d}(\beta)\equiv\frac{1}{n}\sum_{i=1}^{n}d_i(\beta),\ M\equiv l+m$$

さらに，同時パターンの各セルに対するダミー変数を $b_{jk}$，同時パターンの行と列に対するダミー変数をそれぞれ $c_j$，$d_k$ とする。また，各グループに対するダミー変数を $a_\gamma$ とすれば，これらにより，個体 $P_i$ に関するグループによる値 $y_i$ と，同時パターンによる値 $z_i$ を次のように定義するものとする。

$$y_i\equiv\sum_{\gamma=1}^{g}a_\gamma\delta_i(\gamma) \tag{2.1}$$

$$z_i^{(1)}\equiv\sum_{j=1}^{l}\sum_{k=1}^{m}b_{jk}\delta_i(j,k) \tag{2.2}$$

$$z_i^{(2)}\equiv\sum_{j=1}^{l}\sum_{k=1}^{m}(c_j+d_k)\delta_i(j,k) \tag{2.3}$$

ここで，$z_i^{(1)}$　$z_i^{(2)}$ は2つのモデルによる値である。さらに，ダミー・ベクトルとして次の諸量を導入する。

$\boldsymbol{a} \equiv [a_1, a_2, \cdots, a_g]$
$\boldsymbol{b} \equiv [b_{11}, b_{12}, \cdots, b_{1m}, b_{21}, \cdots, b_{2m}, \cdots, b_{l1}, b_{l2}, \cdots, b_{lm}]$
$\boldsymbol{c} \equiv [c_1, c_2, \cdots, c_l, d_1, d_2, \cdots, d_m]$

## 3．3－way データの数量化法

### 3.1　パターン値 $z_i^{(1)}$ の数量化法

数量化の目的から，$y_i$ と $z_i^{(1)}$ の相関が最大となるように，ダミー・ベクトル $\boldsymbol{b}$ を決定すれば，式(2.2)により $z_i^{(1)}$ の数量化が行なえる。この場合の相関は正準相関であるので，数量化の際の規準として，

$$\rho^2 = \frac{(\boldsymbol{b}\,\Sigma_{21}^{(1)}\,\boldsymbol{a}')^2}{(\boldsymbol{b}\,\Sigma_{22}^{(2)}\,\boldsymbol{b}')(\boldsymbol{a}\,\Sigma_{11}\,\boldsymbol{a}')} \tag{3.1}$$

をとることとする。ただし，$\Sigma_{11}$，$\Sigma_{22}^{(1)}$，$\Sigma_{21}^{(1)}$ はそれぞれ次のように定義される分散・共分散行列である。

$\Sigma_{11}$ : $(\gamma, \gamma')$ 要素が $\frac{1}{n}\sum_{i=1}^{n}\{\delta_i(\gamma)-\bar{\delta}(\gamma)\}\{\delta_i(\gamma')-\bar{\delta}(\gamma')\}$
　　である $g \times g$ 行列

$\Sigma_{21}^{(1)}$ : $(\alpha, \gamma)$ 要素が $\frac{1}{n}\sum_{i=1}^{n}\{\delta_i(\gamma)-\bar{\delta}(\gamma)\}\{D_i(\alpha)-\overline{D}(\alpha)\}$
　　である $L \times g$ 行列

$\Sigma_{22}^{(1)}$ : $(\alpha, \alpha')$ 要素が $\frac{1}{n}\sum_{i=1}^{n}\{D_i(\alpha)-\overline{D}(\alpha)\}\{D_i(\alpha')-\overline{D}(\alpha')\}$
　　である $L \times L$ 行列

しかしながら，　$\delta_i(\gamma)$，$D_i(\alpha)$ に $\sum_{i=1}^{n}\sum_{\gamma=1}^{g}\delta_i(\gamma)=n$，$\sum_{i=1}^{n}\sum_{\alpha=1}^{L}D_i(\alpha)=\text{const.}$　なる条件があることから，上記の分散・共分散行列はランク落ちがある。したがって，任意の1要素ずつをとり除く必要がある。ここで，$\delta_i(\gamma)$ と $\boldsymbol{a}$ に関しては $g$ 番目，$D_i(\alpha)$ と $\boldsymbol{b}$ に関しては $L$ 番目の要素を除いたものを～をつけて表すと，規準(3.1)は

$$\tilde{\rho}^2 = \frac{(\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(1)}\,\tilde{\boldsymbol{a}}')}{(\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{22}^{(1)}\,\tilde{\boldsymbol{b}}')(\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{11}\,\tilde{\boldsymbol{a}}')} \tag{3.1$'$}$$

となる。数量化分析II類の本来の規準は相関としているが，計算面での効率を考え，[3] 本研究では規準(3.1)′をとる。式(3.1)′を次の条件

$$\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{22}^{(1)}\,\tilde{\boldsymbol{b}}'=1 \tag{3.2}$$

$$\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{11}\,\tilde{\boldsymbol{a}}'=1 \tag{3.3}$$

のもとで最大化することを考える。Lagrange の未定数 $\lambda$，$\mu$ を用いると，

$$\varphi \equiv (\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(1)}\,\tilde{\boldsymbol{a}}')-\frac{\lambda}{2}(\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{11}\,\tilde{\boldsymbol{a}}'-1)-\frac{\mu}{2}(\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{22}^{(1)}\,\tilde{\boldsymbol{b}}'-1)$$

とおける。これにより，$\partial\varphi/\partial\tilde{\boldsymbol{a}}=0$，$\partial\varphi/\partial\boldsymbol{b}=0$ を求めると

$$\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(1)}-\lambda\,\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{11}=\boldsymbol{0} \tag{3.4}$$

$$\tilde{\Sigma}_{21}^{(1)}\,\tilde{\boldsymbol{a}}'-\mu\,\tilde{\Sigma}_{22}^{(1)}\,\boldsymbol{b}'=\boldsymbol{0} \tag{3.5}$$

となる。(3.4)×$\boldsymbol{a}'$, $\boldsymbol{b}$×(3.5)より

$$\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(1)}\,\tilde{\boldsymbol{a}}'-\lambda\,\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{11}\,\tilde{\boldsymbol{a}}'=0$$

$$\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(1)}\,\tilde{\boldsymbol{a}}'-\mu\,\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{22}\,\tilde{\boldsymbol{b}}'=0$$

を得る。したがって，条件(3.2)，(3.3)を考慮すると $\lambda=\mu$ となる。式(3.4)と(3.5)にこれを代入して

$$\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(1)}-\lambda\,\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{11}=\boldsymbol{0} \tag{3.6}$$

$$\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{12}^{(1)}-\lambda\,\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{22}^{(1)}=\boldsymbol{0} \tag{3.7}$$

であるから，(3.6)×$\lambda$，(3.7)×$\tilde{\Sigma}_{22}^{(1)-1}\cdot\tilde{\Sigma}_{21}^{(1)}$により

$$\lambda\,\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(1)}-\lambda^2\,\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{11}^{(1)}=\boldsymbol{0}$$

$$\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{12}^{(1)}\,\tilde{\Sigma}_{22}^{(1)-1}\,\tilde{\Sigma}_{21}^{(1)}-\lambda\,\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(2)}=\boldsymbol{0}$$

を得る。これらを辺々加えると

$$\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{12}^{(1)}\,\tilde{\Sigma}_{22}^{(1)-1}\,\tilde{\Sigma}_{21}^{(1)}-\lambda^2\,\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{11}=\boldsymbol{0} \tag{3.8}$$

なる一般固有値問題に帰着できる。また，同様の手順から

$$\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{21}^{(1)}\,\tilde{\Sigma}_{11}^{-1}\,\tilde{\Sigma}_{12}^{(1)}-\lambda^2\,\tilde{\boldsymbol{b}}\,\tilde{\Sigma}_{22}^{(1)}=\boldsymbol{0} \tag{3.9}$$

も得られる。しかしながら，一般にダミー変数の数，あるいは式(3.8)の次元は，式(3.9)の次元よりも小さい。したがって，計算上はまず式(3.8)により$\tilde{\boldsymbol{a}}$を求め，式(3.7)と合わせて$\tilde{\boldsymbol{a}}$を

$$\tilde{\boldsymbol{b}}=\frac{1}{\lambda}\,\tilde{\boldsymbol{a}}\,\tilde{\Sigma}_{12}^{(1)}\,\tilde{\Sigma}_{22}^{(1)-1}$$

と求める。これと式(2.2)により，同時パターン $P_h$に対する数量化が

$$z_h^{(1)}=\tilde{\boldsymbol{b}}\,\tilde{\boldsymbol{D}}_h'$$

として得られる。ただし，パターン・ベクトル $\tilde{\boldsymbol{D}}_h\equiv[D_h(1),\ D_h(2),\ \cdots\cdots,\ D_h(L-1)]$ である。

### 3.2　パターン値 $z_i^{(2)}$の数量化法

モデル式(2.2)による数量化の際には，一般に行列 $\Sigma_{22}^{(1)}$ は大型行列となり，実用上計算に困難を生ずる。本節では，データの持つ情報の縮少を図った数量化法を述べる。ここで行なう情報の圧縮とは，個体の2次元パターンを行と列それぞれに関して総和をとることである。したがってダミー・ベクトル $\boldsymbol{c}$ は各行と列の総和に作用する。パターン値 $z_i^{(1)}$ の数量化と同様に，$y_i$ と $z_i^{(2)}$ の正準相関を規準すると

$$\rho^2=\frac{(\boldsymbol{c}\,\Sigma_{21}^{(2)}\,\boldsymbol{a}')^2}{(\boldsymbol{c}\,\Sigma_{22}^{(2)}\,\boldsymbol{c}')(\boldsymbol{a}\,\Sigma_{11}\,\boldsymbol{a}')} \tag{3.10}$$

となる。ただし，$\Sigma_{21}^{(2)}$, $\Sigma_{22}^{(2)}$ はそれぞれ分散・共分散行列で次のように定義される。

$\widetilde{\Sigma}_{21}^{(2)}$ : $(\beta, \alpha)$要素が$\frac{1}{n}\sum_{i=1}^{n}\{\delta_i(\gamma)-\bar{\delta}(\gamma)\}\{d_i(\beta)-\bar{d}(\beta)\}$で

ある$M\times g$行列

$\widetilde{\Sigma}_{22}^{(2)}$ : $(\beta, \beta')$要素が$\frac{1}{n}\sum_{i=1}^{n}\{d_i(\beta)-\bar{d}(\beta)\}\{d_i(\beta')-\bar{d}(\beta')\}$で

ある$M\times M$行列

しかしながら，前節と同様に，$\delta_i(\gamma)$と$d_i(\beta)$に$\sum_{i=1}^{n}\sum_{\gamma=1}^{g}\delta_i(\gamma)=n$, $\sum_{i=1}^{n}\sum_{\beta=1}^{l}d_i(\beta)=\sum_{i=1}^{n}\sum_{\beta=l+1}^{M}d_i(\beta)=$ const. なる条件があることから，$\Sigma_{11}^{(2)}$, $\Sigma_{22}^{(2)}$ はランク落ちがある。そこで，$\delta_i(\gamma)$より任意の要素を1つ，$d_i(\beta)$より$1\leqq\beta\leqq l$, $l+1\leqq\beta\leqq M$ の任意の要素を1つずつ取り除く必要がある。ここでは，$\delta_i(\gamma)$と$\boldsymbol{a}$に関しては$g$番目，$d_i(\beta)$と$\boldsymbol{c}$に関しては$l$番目と$M$番目の要素を除いたものを$\widehat{\ }$をつけて表すと，規準(3.10)は

$$\rho^2=\frac{(\widehat{\boldsymbol{c}}\ \widehat{\Sigma}_{21}^{(2)}\widehat{\boldsymbol{a}'})}{(\widehat{\boldsymbol{c}}\ \widehat{\Sigma}_{22}^{(2)}\ \widehat{\boldsymbol{c}}')(\widehat{\boldsymbol{a}}\ \widehat{\Sigma}_{11}\widehat{\boldsymbol{a}'})}$$

となる。この規準を，$\widehat{\boldsymbol{c}}\ \widehat{\Sigma}_{22}^{(2)}\ \widehat{\boldsymbol{c}}'=1$, $\widehat{\boldsymbol{a}}\ \widehat{\Sigma}_{11}\ \widehat{\boldsymbol{a}'}=1$ なる条件のもとで，最大化する$\hat{c}$を求めると，3.1と同様の計算から

$$\widehat{\boldsymbol{c}}=\frac{1}{\hat{\lambda}}\widehat{\boldsymbol{a}}\widehat{\Sigma}_{12}^{(2)}\widehat{\Sigma}_{22}^{(2)-1}$$

となる。これと式(2.3)より同時パターン$P_h$に対する数量化が

$$z_h^{(2)}=\widehat{\boldsymbol{c}}\,\widehat{\boldsymbol{d}}_h'$$

として与えられる。ただし，$\widehat{\boldsymbol{d}}_h\equiv[d_h(1),\ d_h(2),\ \cdots,\ d_h(l-1),\ d_h(l+1),\ d_h(l+2),\ \cdots,\ d_h(M-1)]$ である。

## 結　言

林の数量化分析II類の3－wayデータへの拡張としての手法を提案した。本手法において，分類情報を，新しい個体の分類の後に与え，成功した都度ダミー変数その他を更新すれば，学習できることになる。また，本手法によれば，文字の物理的特性情報のみならず，パターン・ベクトルに対して機能的特性の成分を付加することにより，双方の情報を用いた文字認識も可能である。データ解析の一手法としても，応用の範囲は広い。

なお，本研究は文部省在外研究長期(乙)並びにフランス政府給費研修の一環として行なわれたものである。この機会を与えて下さった，本学部情報工学専攻の諸先生方，並びに，計算機使用その他にご助力を戴いた，ボルドー第1大学GRAI研究所のG. Doumeingts氏，C. Berard氏さらに株式会社i2sのA.Ricros社長，E.Chapoulaud氏に感謝する。

## 参考文献

1）河口至商：多変量解析入門Ⅰ（1973），森北出版
2）河口至商：多変量解析入門Ⅱ（1978），森北出版
3）田中豊・垂水共之・脇本和昌編：パソコン統計解析ハンドブックⅡ（1984），共立出版
4）提田敏夫：電子通信学会情報・システム部門全国大会論文集（1981），pp, 1－66
5）提田敏夫：電子通信学会総合全国大会論文集（1982），pp, 8－401
6）渡辺洋一・行場次朗・平田　忠・丸山欣哉：テレビジョン学会誌 vol. 39. No. 6 (1986), 509－515
7）新保　勝・佐藤義治・山ノ井髙洋・河口至商：北大工学部研究報告 82（1976），125－132
8）Yamanoi, T., Shimbo, M., Sato, Y., Kawaguchi, M. : TENSOR N. S. 33 (1979) 66. 343－346
9）岩坪秀一：電子技術総合研究所研究報告 801（1983），1－92