数理解析研究所講究録
第820巻 1993年 33-46

確率クラスタリング過程

慶大理工 渋谷政昭 (MASAAKI SIBUYA)
鹿児島大理 大和元 (HAJIME YAMATO)

1. 確率クラスタリング過程 (モデルA)

番号 1, 2, … の付いた玉 $B_1, B_2, \cdots$ (各1個) を, 番号 1, 2, … の付いた壺 $U_1, U_2, \cdots$ にでたらめに入れる. 1個の壺には玉を何個でも入れることができる. 入り方の確率規則を次の通りとする. ただし $\alpha > 0$ は本稿で一貫して用いるパラメータである.

1° B_1 を U_1 に入れる.

2° 以下 $n = 1, 2, \cdots$ について, "玉 $B_1, \cdots, B_n$ が壺 $U_1, \cdots, U_u$ ($1 \le u \le n$) に入っており, U_j に入っている玉の数が v_j ($1 \le v_j$, $\sum_{j=1}^{u} v_j = n$)" の条件下で,

B_{n+1} は U_j に確率 $v_j/(\alpha+n)$ で入り, $1 \le j \le u$,

U_{u+1} に確率 $\alpha/(\alpha+n)$ で入る. □

同じ壺に入る玉をクラスタとみるし, 上記の過程を'確率クラスタリング過程'とよぶ. 本稿でその不変性を考えるが, 上記

の，壺と玉を区別する場合を‘モデルA’と呼ぶ．n個の玉が入った状態を‘n時点’と呼び，本稿では主として$n<\infty$における粉玉諸量の分布を議論する．

玉B_nが入る壺の番号をX_n，$u=\max_{1\le j\le n} X_j$，$A_{nj}=\{i: X_i=j\}$，$1\le j\le u$，$V_{nj}=|A_{nj}|$（集合A_{nj}の基数）とし，

$$A^{(n)}=(A_{n1},\cdots,A_{nu})\quad;\quad V^{(n)}=(V_{n1},\cdots,V_{nu})$$

などについて議論する．$A^{(n)}$の可能な実現値は，

$$\mathcal{A}_n=\{a=(a_1,\cdots,a_u),\ a_j\neq\phi,\ a_j\subset\mathbb{N}_n,\ a_i\wedge a_j=\phi\ (i\neq j),$$
$$\bigcup_{j=1}^{u} a_j=\mathbb{N}_n,\ \min\{k\in\mathbb{N}_n,\ k\notin\bigcup_{i=1}^{j} a_i\}\in a_{j+1}\}$$

という$\mathbb{N}_n=\{1,\cdots,n\}$の，順序ある，制限付きの分割である．また$V^{(n)}$の可能な実現値は

$$\mathcal{V}_n=\{v=(v_1,\cdots,v_u): v_j>0,\ \sum_{j=1}^{u} v_j=n\}$$

という，順序ある，自然数nの分割である．

モデルAの前提条件を改めて書くと，

$$(1)\quad\begin{cases} P(X_1=1)=1, \\ P(X_{n+1}=j\mid V^{(n)}=v)=\begin{cases} v_j/(\alpha+n), & 1\le j\le u, \\ \alpha/(\alpha+n), & j=u+1, \end{cases}\end{cases}$$

ただし$v\in\mathcal{V}_n$，である．

さて，$A^{(n)}$の初等分布が次の通りであることを，帰納法により導くことができる：

命題 1. (1) の条件の下で

(2) $$P(A^{(n)}=a)=\frac{\alpha^u}{\alpha^{[n]}}\prod_{j=1}^{u}(\nu_j-1)!,$$

$a=(a_1,\cdots,a_u)\in\mathcal{A}_n$, $\nu_j=|a_j|$, $1\le u\le n$,

$\alpha^{[n]}=\alpha(\alpha+1)\cdots(\alpha+n-1)$

である. □

$A^{(n)}$ の確率分布が $\{\nu_1,\cdots,\nu_u\}$ だけに依存し, ν_j の順序にも a_j の要素にもよらないことを強調しておく. ただし $a\in\mathcal{A}_n$ の制約より a_j の要素は任意ではない.

命題 2. $c^{(n)}=c(c-1)\cdots(c-n+1)$ とする.

(3) $$P(A_{n1}=a_1)=\frac{\alpha(\nu_1-1)!}{(\alpha+n-1)^{(\nu_1)}},\qquad \nu_1=|a_1|=1,\cdots,n.$$

$$P((A_{n1},A_{n2})=(a_1,a_2))\qquad \nu_j=|a_j|,\ j=1,2,$$

$$=\begin{cases}\dfrac{\alpha^2(\nu_1-1)!(\nu_2-1)!}{(\alpha+n-1)^{(\nu_1+\nu_2)}}, & \nu_j>0,\ \nu_1+\nu_2\le n \text{ のとき},\\ \dfrac{\alpha(n-1)!}{\alpha^{[n]}}, & \nu_1=n,\ \nu_2=0 \text{ のとき},\end{cases}$$

$\cdots$

$$P((A_{n1},\cdots,A_{nk})=(a_1,\cdots,a_k))\qquad \nu_j=|a_j|,\ j=1,\cdots,k,$$

$$=\begin{cases}\dfrac{\alpha^k\prod_{j=1}^{k}(\nu_j-1)!}{(\alpha+n-1)^{(\sum_{j=1}^{k}\nu_j)}}, & \nu_j>0,\ \sum_{j=1}^{k}\nu_j\le n \text{ のとき},\\ \cdots & \\ \dfrac{\alpha^2(\nu_1-1)!(\nu_2-1)!}{\alpha^{[n]}}, & \nu_1>0,\nu_2>0,\ \nu_1+\nu_2=n,\\ & \nu_3=\cdots=\nu_k=0 \text{ のとき},\\ \dfrac{\alpha(n-1)!}{\alpha^{[n]}}, & \nu_1=n,\ \nu_2=\cdots=\nu_k=0 \text{ のとき},\end{cases}$$

ただし a_j は $(a_1, \cdots, a_k, \mathbb{N}_n - \cup_{i=1}^k a_i) \in \mathcal{A}_n$ を満たすものに限る. □

命題2から $A_{n1}, A_{n2}, \cdots$ についての条件付確率分布が得られる. たとえば

(4) $$P(A_{n2} = a_2 \mid A_{n1} = a_1) = \frac{\alpha(v_2 - 1)!}{(\alpha + n - v_1 - 1)^{(v_2)}}, \quad v_2 = 1, \cdots, n - v_1.$$

この右辺は(3)の第1式で n を $n - v_1$ に変えたものである. このようなマルコフ性は $(X_n)_{n=1}^{\infty}$ のマルコフ性を反映したものであり, 一般的には次の命題が成り立つ.

命題3.

$(A_{n1}, \cdots, A_{nk}) = (a_1, \cdots, a_k)$, $\sum_{j=1}^k |a_j| = v_0 < n$ という条件の下での $(A_{n,k+1}, \cdots, A_{nn})$ の確率分布は, $\mathbb{N}_n - \{1, \cdots, v_0\}$ を $\mathbb{N}_{n-v_0}$ に変えれば, $A^{(n-v_0)}$ の確率分布に等しい. □

確率クラスタリング過程に戻り, $(X_n)_{n=1}^{\infty}$ について命題3を考えると, $(X_{n_j} = 1)_{j=1}^{\infty}$ とする部分的部分系列を除いた残りが, $(X_n)_{n=1}^{\infty}$ と同じ構造をもつことになる. さらに $X_{n_j} = 2$ の部分系列, $= 3$ の部分系列, … を除いても同様である.

2. 壺を区別できない場合 (モデルB)

確率クラスタリング退化で，壺を区別できない場合，$A^{(n)}$ は単に $\mathbb{N}_n$ の分割とみなすことができる．$A^{(n)}$ の順序を無視したものを $A^{(n)*}$ で表わそう．その実現値は $\mathbb{N}_n$ の分割の全体

$$\mathcal{A}_n^* = \left\{ \{a_1, \cdots, a_u\} : a_j \neq \phi,\ a_i \cap a_j = \phi,\ i \neq j,\ \bigcup_{j=1}^{u} a_j = \mathbb{N}_n,\ 1 \leq u \leq n \right\}$$

である．部分集合の大きさを表わすにも，順序のあるびを用いるのは不適当で，"順序統計量" $v_{(1)} \leq \cdots \leq v_{(u)}$ を用いるか，あるいはこれと同等であるが，$a_i = k$ となる i (あるいは v_i) の数 s_k, $1 \leq k \leq n$, を用いる．

$$s = (s_1, \cdots, s_m),\ s_k \geq 0,\ \sum_{k=1}^{m} k s_k = n$$

を分割の'大きさの指標'と呼ぶ．s は自然数 n の分割に他ならず，$1^{s_1} 2^{s_2} \cdots n^{s_n}$ という表示法も用いられている．$u = \sum_{k=1}^{n} s_k$ は部分集合の数 (玉の入った壺の数) である．

とはいうものの，空の壺に玉を入れるときに，もっとも小さい番号の壺に入れるために $\mathcal{A}_n$ は特別の構造をしており，分割 $\{a_1, \cdots, a_u\}$ が与えられれば，その順序 (壺の番号) を復元できる：a_1 は 1 の入っている部分集合，a_2 は，a_1 の要素外で最小のものが入っている部分集合，$\cdots$ である．U_1 は命題Zでも見られるように，当然多くの玉が入る

壺である．したがって，別のモデルとするのは必ずしも適切ではないが，壺の番号を無視して球のみを問題にするときを‘モデルB’と呼ぶ．

命題4. (1)の条件の下で，

$$(5)\qquad P(A^{(n)*}=a^*)=\frac{\alpha^u}{\alpha^{[n]}}\prod_{j=1}^{n}((j-1)!)^{s_j},\quad a^*\in\mathcal{A}_n^*,$$

ただし $s=(s_1,\cdots,s_n)$ は分割 a^* の大きさの指標，$u=\sum_{k=1}^{n}s_k$，である．□

(5)は(2)の書き換えにしか過ぎないが，$A^{(n)*}$ の分布が s にしか依存しないことを改めて強調する．n時点での玉の番号を $B_{\nu_1},\cdots,B_{\nu_n}$ ($(\nu_1,\cdots,\nu_n)$ は $\mathbb{N}_n$ の任意の順列)に変えても(5)には関係しない．つまり，(5)は要素の置換に関して不変である．

この事実から，たとえば，“任意の $i,j\in\mathbb{N}_n$ が同じ部分集合に入る確率”は“1,2が同じ部分集合に入る確率”に等しく，したがって $A^{(n)}$ の分布について計算する必要はなく，

$$P(A^{(2)}=(\{1,2\}))=\frac{1}{\alpha+1}$$

と求まる．同様の考察で，いろいろな事象の確率が計算できるが，一般的には次のようにまとめることができる．

$\mathbb{N}_n=E_1\cup E_2$，$E_1\cap E_2=\phi$，$|E_1|=m$，$0<m<n$

とし，$A^{n(*)} \cap E_i$，$i=1,2$，で $\{A_{n1} \cap E_i, \cdots, A_{nu} \cap E_i\}$ から空集合を除いたものとする．

命題 5. b_i^*，$i=1,2$，を E_i の任意の細分割とする．

(i) $a^* = b_1^* \cup b_2^*$ のときに

$$P(A^{(n)*} = a^* \mid A^{(n)*} \cap E_2 = b_2^*)$$
$$= P(A^{(n)*} \cap E_1 = b_1^* \mid A^{(n)*} \cap E_2 = b_2^*) = P(A^{(m)*} = b_1^{**})$$

ただし b_1^{**} は b_1^* の要素を $\mathbb{N}_m$ の要素に変えたものである．

(ii) 上の (i) と同じ記号で，

$$P(A^{(n)*} \cap E_1 = b_1^*) = P(A^{(m)*} = b_1^{**}).$$

左辺は $A^{(n)*}$ の任意の m 要素を取り上げ，他を無視したときの部分分割の分布である．□

これらの事実は部分クラスタリング過程に関して本質的であり，逆に分布 (2) を次のように特徴づけることができる．

命題 6. 任意の時点 n における $\mathcal{A}_n$ 上の部分分布 $P(A^{(n)} = a)$，$a \in \mathcal{A}_n$，が…

(i) a の成分の基数の集合 $\{|a_1|, \cdots, |a_u|\}$ だけに依存し，

(ii) $A_{n1} = \{1, \cdots, k\}$，$1 \le k \le n-1$，の条件の下での $(A_{n2}, \cdots, A_{nu})$ の部分分布が，要素 $\{k+1, \cdots, n\}$ を $\mathbb{N}_{n-k}$ に変えれば，$A^{(n-k)}$ の部分分布に等しい．

このとき，$P(A^{(n)}=a)$ は (2) 式となる．□

命題 7.

命題 6 の条件 (ii) を次のように変えることができる：

(ii′) $\mathbb{N}_m$ に属する任意の要素を $A^{(n)}$ から除いたときの確率分布 が，要素を ((たとえば 1)) $\mathbb{N}_{m-1}$ に変えたとき，$A^{(n-1)}$ の分布に等しい．□

3. 玉を区別できない場合（モデル C）

確率クラスタリング過程において玉が区別できず，壺に入る玉の数だけが観測可能であるとする．時点 n で $U_1, \cdots U_u$ に入る玉の数を $V^{(n)}=(V_{n1}, \cdots, V_{nu})$ とする．

命題 8. (1) の条件の下で，

$V^{(n)}=(V_{n1}, \cdots, V_{nu})$，$V_{nj}=|\{i: X_i=j\}|$，の確率分布は，

$$(6)\quad P(V^{(n)}=v)=\frac{\alpha^u}{\alpha^{[n]}}\frac{(n-1)!}{\prod_{j=1}^{u-1}(n-\sum_{i=1}^{j}v_i)},$$

$v=(v_1, \cdots, v_u)\in \mathcal{V}_m$，$u=1, \cdots n$．□

この場合を“モデル C”と呼ぶことにする．モデル A で単に v とする $a\in\mathcal{A}_m$ の数をかぞえればよい．逆に，モデル C で v_1 個の玉に，1 と $\mathbb{N}_m\setminus 1$ から選んでランダムな番号を付け，残りに逐次 $a\in\mathcal{A}_m$ の条件を満た

すようにランダムに付ければモデル A を得る.

モデル A の命題 2 に対応する $V^{(n)}$ の周辺分布は次の通りである.

命題 9.

(7) $$P(V_{n1}=v)=\frac{(n-1)^{(v-1)}\alpha^{[n-v]}}{(\alpha+1)^{[n-1]}},\qquad v=1,\cdots,n$$

あるいは

$$P(V_{n1}-1=x)=\binom{\alpha+n-x-2}{n-x-1}\Big/\binom{\alpha+n-1}{n-1}$$

$$=\binom{-1}{x}\binom{-\alpha}{n-1-x}\Big/\binom{-\alpha-1}{n-1},\qquad x=0,\cdots,n-1$$

これは負の超幾何分布 $NgHg(n-1;\ 1,\alpha)$, つまり 2 項分布 $Bn(n-1,p)$ の p がベータ分布 $Be(1,\alpha)$ に従うベータ・2項分布である. □

周辺分布 $P((V_{n1},V_{n2})=(v_1,v_2))$ なども同様の形となる. 条件付分布も

$$P(V_{n2}=v_2\mid V_{n1}=v_1)=\frac{(n-v_1-1)^{(v_2-1)}\alpha^{[n-v_1-v_2]}}{(\alpha+1)^{[n-v_1-1]}},$$

$$v_2=1,\cdots,n-v_1.$$

これは, (7) の n を $n-v_1$ に変えたものとなっている. 一般に,

命題 10. $(V_{n1},\cdots,V_{nk})=(v_1,\cdots,v_k)$, $\sum_{j=1}^{k}v_j=v_0<n$ という条件の下での $(V_{n,k+1},\cdots,V_{nu})$ の確率分布は,

(6) で n を $n-v_k$ に変えたものである. □

これは命題5に相当するものである. 命題6に相当するものは次の通りである.

命題11. $V^{(n)}$ が (6) の確率分布に従うとする. n 個の玉のうちの1個をランダムに選んで除くとする. もしも $v_k=1$ の壺 U_k を選ぶと空の壺が生じるが, そのときには $U_{k+1}, \cdots, U_n$ の番号をひとつずつ前にずらす. このときの玉の数 $V^{(n-1)}$ の分布は (6) で n を $n-1$ に変えたものとする. □

後述のように $n\to\infty$ の極限を考えるときにはモデルCが基本的となる.

4. 玉も壺も区別できる場合 (モデルD)

最後に, 玉も壺も区別できず, 部分集合の大きさの指標でのみ観測できる場合を'モデルD'とする. 確率的な指標を $S^{(n)}=(S_{n1},\cdots,S_{nn})$ とする. その実現値の全体は自然数 n の分割の全体 $\mathcal{J}_n$ である.

命題12. 確率クラスタリング過程の n 時点で, 分割の大きさの指標 $S^{(n)}$ の確率分布は

$$P(S^{(n)}=s)=\frac{\alpha^u}{\alpha^{[n]}}\,\frac{n!}{\prod_{j=1}^{n}(j^{s_j}s_j!)}, \tag{8}$$

$$s=(s_1,\cdots,s_m)\in\mathcal{S}_n\ ,\quad u=\sum_{j=1}^{m}s_j,$$

である. □

分布 (8) はモデルBからは, 多項係数を用いた計算で直ちに得られるが, モデルCから導くことはそれほど単純ではない. 逆にモデルDからモデルBを導くのも, 単にうさぎんる番号付けを与えるだけであるが, モデルCを導くには技巧が必要である.

命題13. $S^{(n)}$ が (8) の分布に従っているとする. n個の要素のうちの1つを等確率で選び, それが属している部分集合(基数を V_1 とする)を除く, 残りの $n-V_1$ 個の要素の1つを等確率で選び, それが属している部分集合(基数を V_2 とする)を除く. ・・・ このとき $(V_1, V_2, \cdots, V_n)$ はモデルCの分布 (6) に従う. □

したがってn個の要素の1つを等確率で選ぶ. その要素が属している部分集合全体を除いても, その要素1個を除いても (8) でnから除いた数だけ減らしたものが得られる.

5. 4つのモデル

4つのモデルA－Dを適当に使い分けることにより確率の計算が容易となる. 本質的な特徴は, 正の番号の

交換不変性（exchangeability）である．それぞれの確率分布を下の表にまとめておく．

確率クラスタリング過程の諸分布

垂 \ 玉	区別あり	区別なし
区別あり	A: $P(A^{(n)}=a)=\frac{\alpha^u}{\alpha^{[n]}}\prod_{j=1}^{u}(v_j-1)!$	C: $P(V^{(n)}=v)=\frac{\alpha^u}{\alpha^{[n]}}\frac{(n-1)!}{\prod_{j=1}^{u-1}(n-\sum_{i=1}^{j}v_i)}$
区別なし	B: $P(A^{(n)*}=a^*)=\frac{\alpha^u}{\alpha^{[n]}}\prod_{j=1}^{n}((j-1)!)^{s_j}$	D: $P(S^{(n)}=s)=\frac{\alpha^u}{\alpha^{[n]}}\frac{n!}{\prod_{j=1}^{n}(j^{s_j}s_j!)}$

A: $a=(a_1,\cdots,a_u)\in\mathcal{A}_n$．$a_j$：垂 U_j に入っている玉の番号の集合．
$v_j=|a_j|>0$．

B: $a^*=\{a_1,\cdots,a_u\}\in\mathcal{A}_n^*$：$\mathbb{N}_n$ の分割の全体
$s=(s_1,\cdots,s_n)\in\mathcal{J}_n$：$n$ の分割の全体

C: $v=(v_1,\cdots,v_u)$　　Aと同じ

D: $s=(s_1,\cdots,s_n)$　　Bと同じ

いずれも（パラメータ）α を含む部分は同じ形であり，玉の入っている垂の数 u は十分統計量である．その確率分布は，

$$(9)\qquad P(U_n=u)=\begin{bmatrix}n\\u\end{bmatrix}\frac{\alpha^u}{\alpha^{[n]}},\quad u=1,\cdots,n,$$

ただし $\begin{bmatrix}n\\u\end{bmatrix}$ は第1種スターリング数（Knuthの記号）

であって，等式 $\alpha^{[n]} = \sum_{u=1}^{n} \begin{bmatrix} n \\ u \end{bmatrix} \alpha^u$ によって定まる．

6. モデル C の極限分布．

モデル C において $V^{(n)}/n = (V_{n1}/n, \cdots, V_{nu}/n)$ とし，$n \to \infty$ とすると，無限次元単体上の確率分布が得られる．負の超幾何分布に従う V_{n1} をnで割った極限は，2項分布の比率が確率に収束するから，ベータ分布であり，今の場合，その確率密度関数は

(10) $$\alpha(1-u)^{\alpha-1} \cdot 1[0<u<1]$$

である．（$1[\cdot]$ は平均の定義関数．）$V^{(n)}$ のマルコフ性（命題13）より，極限も同じ性質をもつ．

命題14.

$(W_n)_{n=1}^{\infty}$ を (10) に従う (0,1) 上の（i.i.d. 確率変数の系列とし，$(U_n)_{n=1}^{\infty}$ を

$$U_1 = W_1, \quad U_n = W_n \prod_{j=1}^{n-1} (1-W_j), \quad n=2,\cdots$$

で定義する．$V^{(n)}/n$ $(n \to \infty)$ は $(U_n)_{n=1}^{\infty}$ に収束する．さらに $(X_1, \cdots, X_n)$ は $(U_n)_{n=1}^{\infty}$ からの確率標本である．□

$(U_n)_{n=1}^{\infty}$ は確率クラスターが過程の諸性質を継承している．

7. おわりに.

本稿は独創的なものではなく，最近集団遺伝学で得られている成果を，集団遺伝学の用語・概念を使わずに解説したものである．ただし p.12 の表のような 2×2 のモデルとして整理することにより体系的に，明快となったと自負している．文献は非常に多いが次の2つの総合報告に詳細な文献表がある．

- Hoppe, F.M. (1987) The sampling theory of neutral alleles and an urn model in population genetics, J. Math. Biology 25 123-159.
- Ewens, W.J. (1990) Population genetics theory — The past and the future, S. Lessard (ed.) Mathematical and Statistical Developments of Evolutionary Theory, 177-227, NATO Adv. Sci. Inst. Ser. C 299, Kluwer, Dordrecht, Holland.