〔生物工学会誌　第85巻　第10号　446–449. 2007〕

講座 ケモメトリックス入門

第7回　その他のケモメトリックスの方法

錦織　理華*・川瀬　雅也

これまで統計学の分類に従い，種々の解析法の基本を見てきた．最終回に当たる今回は，最近いろいろな分野において使われ始めている自己組織化マップ（self organizing map；SOM）と，ベイズ統計について解説を行う．本連載では取り上げることのできなかった，あるいは十分に説明の尽くせなかった事項も多々あり，これについては別の機会に，是非本誌上で解説を行いたいと考えている．

SOM

SOMは本誌84巻6号のバイオミディア（p.449）において，越智が解説しているように，1980年代にKohonenにより提唱され，高次元の入力データの類似関係を主として2次元のマップ上に投影する教師なし分類手法として発展し，広く利用されている．DNAマイクロアレーの解析，メタボロミクスデータの解析，あるいは社会科学的なデータ解析など，利用分野を挙げればきりがないくらいである．

まず，簡単にSOMの原理を解説する．上述の越智の解説と合わせて見て頂ければ分かりやすくなると思う．N個のM次元データ，$x_k = \{x_1, x_2, \cdots x_M\}$ $(k = 1, \cdots N)$ の各データの類似について検討する場合を考える．はじめに，データを投影する対象となるマップを準備する．マップは$d \times l$の格子状になっており，格子の数はデータ数と用途に応じて解析者が決める必要がある．

このマップの各格子点の上には，解析するデータと同じ次元の参照ベクトル $m_i = \{m_1, m_2, \cdots m_M\}$ が準備される，各m_iの格子点状の位置を表す2次元ベクトルとして$r_i = (s, t)(s = 1 \cdots d, t = 1 \cdots l)$を定義する．

この最初の時点では，各参照ベクトルのデータは乱数で構成されている．このマップを，データに対応した形に学習させていくのが次の段階である．

この段階では，以下のステップが繰り返される．繰り返しの回数を$t = 1, 2, \cdots, T$と置くとt回目の計算では，

1）1つの入力データx_kをこのマップ上のすべての参照ベクトルm_i^tと比較し，そのユークリッド距離$\| x - m_i^t \|$がもっとも小さくなる格子点を決定する．この格子点をcとすると，$\| x_k - m_c^t \| = min\{\| x_k - m_i^t \|\}$

2）格子点cを中心とし，マップ上の距離が近い格子点の参照ベクトルmが，データx_kと近い値になるように，以下の式を用いて学習させる．

$$m_i^{t+1} = m_i^t + h_{ci}^t [x - m_i^t]$$

3）このh_{ci}^tは近傍関数と呼ばれ，格子間のマップ上の距離$\| r_c - r_i \|$と，計算回数tで定義される．$\| r_c - r_i \|$が大きいcから離れた格子ではh_{ci}^tが0になり学習が行われない．また計算回数が大きくなるとh_{ci}^tが0になり計算が収束する．h_{ci}^tにはいくつかの式が提案されており，そのうち2つを紹介すると

① cからの位置が一定の範囲N_cに含まれる格子の参照ベクトルに同程度学習させる方法で，

$h_{ci}^t = \alpha(t)$　$i \in N_c^t$，$h_{ci}^t = 0$　$i \notin N_c^t$ $(0 < \alpha(t) < 1)$

となる．$\alpha(t)$はtの増加とともに単調減少し0に近づく．N_c^tの範囲も同様に小さくなる．

②比較的汎用されている関数で，ガウス関数を用いた以下の式で定義される．

$$h_{cj}^t = \alpha(t) * \exp\left(\frac{-\|rc - ri\|^2}{2R(t)^2}\right)$$

$\alpha(t)$は学習率と呼ばれる値で，$R(t)$は上記のN_c^tと同様に学習させる範囲を定義する値である．$\alpha(t)$および$R(t)$は回数tの増加に応じて単調に減少する．

4）上記1），2）の手順をn個のデータxについて行い，1サイクルとする．

以上の手順で，構築されたマップの参照ベクトルm_i^Tと各データを比較し，$\| x - m_i^T \|$がもっとも小さくなる格子点上にデータをプロットしていくと，類似のデータが近くの格子点に集まった結果のマップが得られる．

*著者紹介　大阪大谷大学薬学部（助教）　E-mail: nisikir@osaka-ohtani.ac.jp

SOMの利用先としては，DNAマイクロアレー解析や疾患発生に関わる遺伝子の探索などに利用されている．

ちょっと変わった統計

確率　高校の数学を思い出していただきたい．「確率・統計」という科目になっていたと思う．「確率」と「統計」は，実は切っても切れない密接な間柄なのである．

それでは確率と統計の関係について考えてみよう．正常なサイコロを考えてみる．このサイコロを振ったとき，どの目の出る確率も$\frac{1}{6}$であると，誰でも考えるが本当であろうか．試しに6回振ってみよう．どの目も1回ずつ出ることなど滅多にないことがすぐにわかる．では，なぜ，どの目の出る確率も$\frac{1}{6}$であると考えるのであろうか．これは，サイコロを何万回も振って各々の目の出る回数を数えていくと，どの目の出る回数もほぼ等しくなり，結果として，どの目の出る確率も$\frac{1}{6}$であると結論付けることができるという「相対頻度の極限による確率の定義」に基づいているからである．別の考え方として，最初から，どの目の出る確率も等しいと考えることから出発する考え方がある．こう考えると，自然と，どの目の出る確率も$\frac{1}{6}$であるという結論が導かれる．この考え方を「等確率性による確率の定義」という．どちらの考え方に従うにせよ，確率計算の前提として，ある事象の起こる確率の分布，つまり確率分布を前提として持っている．

これまで見てきた計量薬学手法（推計統計学的手法）をもう一度見直してみると，“「母集団の分布」が「正規分布」に従う”という表現が度々出てきたかと思う．では，この母集団の分布が正規分布に従うとは何を根拠に決められているのか．推計統計学では，ある事象を確認するための検証（生物工学分野では実験であることが多い）を非常に多くの回数を繰り返して行い，最終的に得られるその事象の起こる頻度確率を基に理論の組み立てが行われている．この頻度確立の分布が，推計統計学の理論を形成する各種の統計分布となってくるわけである．つまり統計学は，その基礎として“確率分布”の概念を持っているということを忘れてはならないのである．

推計統計学でよく使われる確率分布　本項までに，既出の確率分布としては“正規分布”，“χ^2分布”，“F分布”，“t-分布”がある．この他に，“β分布”，“Γ分布”，“指数分布”，“対数正規分布”などがある．これらの分布は，確率変数の値が連続量であるので“連続分布”と呼ばれている．確率変数とは，事象を意味していると考えていただければいい．先のサイコロの例で言えば，「1の目が出る」という事象は一つの確率変数となる．一方，確率変数の値が離散量となる分布もある．このような分布には“二項分布”，“ポアソン分布”，“超幾何分布”などがある．これらの分布も推計統計学で用いられる．

上記分布の中で，医学・薬学分野で比較的よく使われる対数正規分布について説明を行う．対数正規分布は，確率変数の対数を取ったときに，その分布が正規分布に従うというものである．形は，“χ^2分布”や“F分布”に似たものとなっている．対数正規分布の理論式は次のようなもので，

$$f(x)=\begin{cases}\dfrac{1}{\sqrt{2\pi\sigma x}}\exp\left\{\dfrac{-(\log x-\mu)^2}{2\sigma^2}\right\} & ;x>0\\ 0 & ;x\le 0\end{cases}$$

血清タンパク質やクレアチニンなどの血液性化学検査の値などがこの分布に従うことが知られている．

本書では他の個々の分布について説明しないが，ここに挙げた確率分布がどのようなものか，興味のある方は是非インターネットで検索してみてほしい．初めて見るような分布の中に，自分の系を説明するのに必要な分布が見いだせるかもしれない．

多くの統計学の書籍では，解析手法に重点が置かれるあまり，その背景にある確率分布に注意を促す記述はあまり見受けられない．しかし，実際の解析を行うにあたっては，確率分布が大切な要因となることを十分に理解していただきたい．

確率分布を十分に利用して理論を展開している統計学にベイズ統計学（ベイズ推計学とも呼ばれる）がある．次に，ベイズ統計学の簡単な説明を行う．

ベイズ統計学　ベイズ統計学は18世紀にイギリスの牧師であり数学者であるThomas Bayseが提案した統計学である．この統計学では，事象の起こる確率分布を，解析を行う個人の経験などを基に決めて，それに基づき解析を行う手法である．個人の経験・主観に基づくということで，事前に決められる事象の発生確率を「主観確率」と呼んでいる．一言でベイズ統計学を表せば，「主観確率に基づく，事象の発生の事前分布を仮定した統計学」ということになる．ベイズ統計学では，観測変量もパラメータもすべて確率変数として扱われる．

ベイズ統計学の基礎を説明する．ある事象Bが起こったときに，事象Aが起こる確率をAが起こる条件付確率と呼びP(A|B)で表す．

$$P(A\,|\,B)=\frac{P(A\cap B)}{P(B)}$$

となる．分子はAとBが同時に起こる確率（同時確率）であり，分母はBの起こる確率（周辺確率）である．定義により，

$$P(A|B)P(B)=P(A\cap B)=P(B|A)P(A)$$

であるので，あらゆるAについて考えると，

$$P(A|B)=\frac{P(B|A)P(A)}{\sum_{A}P(B|A)P(A)}$$

となる．これを「ベイズの公式」あるいは「ベイズの定理」と呼ぶ．ベイズの定理をもう少し一般化してみる．パラメータθの事前分布（事前確率密度関数）を$\pi(\theta)$，θの与えられたときのデータxの分布（確率密度関数）を尤度関数と呼び$f(x|\theta)$で表す．θの事後分布（事後確率密度関数）は

$$P(\theta|x)=\frac{f(x|\theta)x(\theta)}{m(x)}\propto f(x|\theta)\pi(\theta)$$

となる．ここでm(x)はxの周辺確率（この場合，周辺密度と呼ぶほうが正しい）であるが，解析的に求めることができない場合が多く，ほとんどの場合省略される．また，$f(x|\theta)$は尤度関数であるので，全範囲で積分しても1とならないことが多く$l(\theta|x)$とあらわされる場合もある．よってベイズの定理の一般的表現は

$$P(\theta|x)\propto l(\theta|x)\pi(\theta)$$

とされる場合がある．

ベイズ統計の一つの例として健康診断を考えてみたい．ある病気の検査を行うとする．利用される検査法は

○病気である場合，98%の確率で陽性となる

○病気でなければ，陽性となる確率は5%である

○日本人のこの病気の罹患率は2%である

ある人がこの検査で陽性となったとする．この人が病気である確率はいくらか（厳密に言えば期待値であるが，分かりやすいので以後も確率という）．少し考えてみていただきたい．単純に98%と答える人も多いかと思うが，少し短絡的ではないか．98%はあくまでも病気である場合の陽性となる確率で，病気かどうか分からない場合は，その確率は98%ではないはずである．

ここでベイズの定理を用いると

$$P=\frac{0.98\times0.02}{0.98\times0.02+0.05\times0.98}=0.286$$

となり，実は検査で陽性となっても，病気である確立は約29%しかないのである．つまり前提条件により，その確率が変わってくる．これがベイズ統計が現在注目されている点である．

ベイズ統計学は，すべてベイズの定理によって理論が展開されるが，これ以上の詳しい内容は本連載の範囲を超えるので，最後に挙げる参考書をご覧いただきたい．

ベイズの理論では，繰り返しになるが，解析者の経験，たとえば医師の経験や薬剤師の経験を事前分布として取り込めるため，生命科学領域では非常に使いやすい統計学であると考えられる．

生命科学領域においては，複数のパラメータを取り扱うのが常である．ベイズ統計学の応用では，これらのパラメータの事後分布からパラメータの推定を行うわけであるが，このためには多変量確率密度関数の多重積分という厄介な問題が生じてくる．そこで，この問題を解決する方法としてマルコフ連鎖モンテカルロ法という手法が用いられる．マルコフ連鎖モンテカルロ法は多変量事後分布より確率標本を発生させベイズの定理につなぐ手法である．この手法の代表的なアルゴリズムとしてギブス・サンプラーアルゴリズムがある．この手法については，モンテカルロ法などの計算物理手法の基礎も必要となり，本書の範囲を超えるため，詳しい解説は控えることにする．最後の参考書に詳しくあるので，そちらをご覧いただきたい．

最後に強調しておきたいのは，従来の推計統計の枠組みだけではなく，ベイズ統計学のような別の枠組みもあり，どれを用いるのがよりよいかを十分吟味する必要があるということである．

参考書

1) 伊庭幸人：ベイズ統計と統計物理，岩波書店（2003）．
2) 渡部　洋：ベイズ統計学入門，福村出版（1999）．
3) 牧　厚志ら：経済・経営のための統計学，有斐閣（2005）．

【付録】ケモメトリックスのソフトについて

これまで種々の手法についての解説が行われ，実際に使ってみたい手法もあったかと思う．しかし，市販の統計ソフトにない手法も多くあり，自分でプログラムを組むだけの時間もないという方に，フリーでありながら本書で紹介されている手法の多くを実行することが可能なRというソフトウェアがある．

Rについての詳しいことは次の参考書をご覧いただきたい．また，すぐにRについて知りたい方は，

http://cse.naro.affrc.go.jp/takezawa/r-tips/r2.html

をご覧いただきたい．

Rの参考書

1) 船尾暢男，The R tips，九天社（2005）．
2) 岡田昌司，The R book，九天社（2004）．

この2冊には，R収録のCD-ROMもついており使い勝手がよい．