確率的コンプレキシティと学習理論

山西　健司

1. はじめに

本稿では、次の2つの統計的推論の問題を扱う。

1. **確率モデルの最適選択**、
2. **逐次的確率予測**。

前者は、データが与えられたとき、これを発生させている確率分布として最もふさわしいモデルを選択する問題である。後者は、データが逐次的に与えられるとき、オンラインで未来のデータの確率分布を予測するという問題である。これらの問題は、統計的推論の基本問題であると同時に、近年発展している確率的規則の計算論的学習理論[6]·[12]·[16]·[10]といった分野の中心的話題である。本稿では、「学習」という言葉は上の2つの統計的推論の問題を意味するものとする。

1では、未知のデータ発生分布に出来るだけ近いモデルを少ないデータ数で選択するためのアルゴリズムが必要になる。また、2では予測誤差の累積が出来るだけ小さくなるようなアルゴリズムが必要となる。このようなアルゴリズムはどのようにして設計できるのか？本稿は「確率的コンプレキシティ」という概念を軸にして、上の問に対する統一的な解答指針が与えられることを示すものである。

確率的コンプレキシティはJ. Rissanenによって提唱された新しい情報量概念であり、大雑把に言って、データ系列を与えられた確率モデルのクラスを用いて符号化する際の最短符号長として定義される。実は上の2つの問題に有効なアルゴリズムの設計は、確率的コンプレキシティを最良に近似するための符号化を設計することに帰着されるのである。以下、確率的コンプレキシティの概念が最適な学習アルゴリズムの設計と解析に本質的な役割を果たす事情を、近年の情報理論と学習理論の結果をふまえて解説する。

やまにし　けんじ　NEC C&C研究所
〒216 川崎市宮前区宮崎 4-1-1

2. 確率的コンプレキシティ

本節では、情報理論における「符号化」という概念が統計学における「確率分布」と見方の違った同一の概念であることを示し、この関係に基づいて確率的コンプレキシティを導入する。

(情報源)**符号化**とはデータ系列を2進系列に変換することである。データ系列 $\mathbf{Y} = Y_1 \cdots Y_m$ は有限アルファベット $\mathcal{A}$ の直積空間 $\mathcal{A}^m$ の元であるとし、$\{0,1\}^*$ を有限長の2進列の集合として、符号化を表す写像を $\phi : \mathcal{A}^m \to \{0,1\}^*$ で表す。ϕ は1対1写像とする。

さらにϕとしては、符号語の系列$\phi(\mathbf{X})\phi(\mathbf{Y})\phi(\mathbf{Z})\cdots$から $\mathbf{X},\mathbf{Y},\mathbf{Z},\cdots$ に対応する符号語をコンマなどの特別な区切り記号などを用いなくとも、正確にその順に分離して復号化できるような符号化を考える。このような性質をもつためには「任意の2つのデータ系列 $\mathbf{X},\mathbf{Y} \in \mathcal{A}^m$ に対して、符号語$\phi(\mathbf{X})$ と$\phi(\mathbf{Y})$ の一方が他の先頭部分に一致することはない」という条件をみたすことが十分である[3]。この条件をみたす符号化を**語頭符号化**とよぶ。我々は符号化のコストを符号語の長さ-「符号長」-で測り、符号長が出来るだけ短くなるような語頭符号化を設計したい。

実は、語頭符号化の条件と符号長とは密接に結び付いている。実際、ϕが語頭符号化であるための必要十分条件は、各$\mathbf{Y}$に対する符号長を$l(\mathbf{Y})$として、次式(Kraftの不等式)をみたすことである[3]。

$$\sum_{\mathbf{Y}\in\mathcal{A}^m} 2^{-l(\mathbf{Y})} \leq 1.$$

一方、各 $\mathbf{Y} \in \mathcal{A}^m$に対して $Q(\mathbf{Y}) \geq 0$ かつ $\sum_{\mathbf{Y}\in\mathcal{A}^m} Q(\mathbf{Y}) \leq 1$ を満たす $Q(\mathbf{Y})$ を $\mathcal{A}^m$上の**劣確率分布**とよぶ。語頭符号ϕが1つ与えられたら、その符号長$l(\mathbf{Y})$に対して、

$$l(\mathbf{Y}) = -\log Q(\mathbf{Y}) \tag{1}$$

によって劣確率分布$Q(\mathbf{Y})$が定義できる(logは底が2の対数を表すものとする)。逆に劣確率分布$Q(\mathbf{Y})$が

1つ与えられると、上の関係によって$l(\mathbf{Y})$を符号長関数とする語頭符号化が存在する(例えば、Shannon-Fano-Elias 符号 [3])。以下、このような符号化を **分布Qに対する符号化**と呼び、(1)を$\mathbf{Y}$のQに対する**Shannon 情報量**とよぶ(本稿では、簡単のため、符号長は非整数値をとることを許すものとする)。このように、(劣)確率分布と語頭符号化は表裏一体の関係にある。

では、符号長を出来るだけ短くするにはどのような確率分布に対して符号化すればよいのか?データを発生させる確率分布P^*(これを**真の分布**とよぶ)がわかっていれば、P^*に対する符号化は平均の意味で最小の符号長をもつことが容易に確かめられる [3]。ところが現実には真の分布P^*は未知である場合が多い。その場合、平均符号長を出来るだけ短くするように符号化するにはどうしたらよいか? 1つの解決策は、たった1つの(劣)確率分布の代わりに、真の分布を含むと思われる(含まなくても良い)**確率モデルのクラス**を1つ導入し、これに対して符号化することである。ここでいう確率モデルとは、何らかの数学的制約の入った確率分布のことを指す。

ところで、「"クラス"に対してデータ系列を符号化する」とはどういうことか?そのクラスの中から最適な確率モデルを1つ選んでそれに対して符号化するという方法もあるだろうし、クラスに属するモデル全体を重み付き平均して符号化するという方法もあるだろう。実に様々な符号化が考えられるのである。

今、$\mathcal{H}$を確率モデルのクラスとし、与えられたデータ系列$Y^m = Y_1 \cdots Y_m$の$\mathcal{H}$に対する**確率的コンプレキシティ(Stochastic Complexity)**[11]を「Y^mをクラス$\mathcal{H}$に対して語頭符号化するときの最小符号長」として定義し、$SC(Y^m : \mathcal{H})$とかく。ここで、最小は$\mathcal{H}$に対するあらゆる符号化に関してとるものとする。Shannon 情報量 (1) は1つの分布に対する符号長であったから、$SC(Y^m : \mathcal{H})$はShannon情報量の一般化と見なすことが出来る。(注意:3.2節の最後により数学的に正確な$SC(Y^m : \mathcal{H})$の定義を与える。)

$SC(Y^m : \mathcal{H})$の正確な値は、通常簡単に計算できるとは限らない。そこで実際は特定の符号化を選んで$SC(Y^m : \mathcal{H})$を近似的に評価することになる。符号化の多様性に対応して、確率的コンプレキシティの近似方法は幾通りも考えられる。どの符号化を選ぶかはまさに、統計的推論の状況に依存するのである。以下、これを具体的な符号化を例に見て行こう。

3. 非逐次的符号化とモデル選択

3.1 2段階符号化

本節では、データ系列$Y^m = Y_1 \cdots Y_m$が一括与えられた時にこれを符号化する方法(これを**非逐次的符号化**とよぶ)及び、これによる確率的コンプレキシティの近似について考える。先ず、$\mathcal{H}$を用いてY^mを以下の2つのステップを踏んで語頭符号化することを考える。(1)$\mathcal{H}$の中から確率モデルを1つ選択し、(2)これを用いて確率モデルと一緒にデータの符号化を行う。このような符号化を**2段階符号化 (Two-part Coding)**[9]とよぶ。2段階符号化に必要な全符号長は「選ばれた確率モデルに対するデータの符号長」と「その確率モデル自身の符号長」の総和として計算できる。選ばれた確率モデルをPとすると、これに対するデータの符号長は$-\log P(Y^m)$で求められる。P自身の記述長はKraftの不等式を満たす符号長関数ℓを1つ固定して$\ell(P)$で計算する。よって、全符号長は

$$-\log P(Y^m) + \ell(P) \tag{2}$$

として計算できる。そこで、(2)を$\mathcal{H}$上のPに関して最小化して得られる量は、2段階符号化による確率的コンプレキシティ$SC(Y^m : \mathcal{H})$の近似と見なすことが出来る:

$$SC(Y^m : \mathcal{H}) \approx \min_{P \in \mathcal{H}} \{-\log P(Y^m) + \ell(P)\}.$$

ここで左辺の量のminimumを達成するようなPは、「与えられたデータ系列Y^mを、2段階符号化によって最も短く符号化出来るような確率モデル」である。このようなモデルこそがデータ生成源の最良のモデルであると見なす、確率モデル選択基準を**MDL (Minimum Description Length) 原理** [8]·[9] とよぶ。MDL 原理では、(2) の量の大小を確率モデルの評価値とし、この値が小さい程、データ発生源をうまく表現していると見なす。以上のように、2段階符号化による確率的コンプレキシティの最良近似を考えることにより、最適モデル選択の戦略が得られる。

3.2 最尤符号化

2段階符号化において、もし$\mathcal{H}$が有限のクラスならば、(2)の左辺の計算は特に問題はない。ところが、$\mathcal{H}$が連続の実数値パラメータで指定されている場合はそれは自明に計算できない。なぜならば、通常、実数

値パラメータの指定には無限の精度を要求されるので、まともに計算すればその符号長が無限大になってしまうからである。この問題を克服する1つの方法として、パラメータ空間の**量子化**に基づく2段階符号化の方法が考えられる[9]·[19]。しかしここでは、確率的コンプレキシティを近似する非逐次的符号化として、最尤符号化とよばれる、より単純な符号化方法を紹介する。

今、$\mathcal{H}^{(k)} = \{P_{\theta,k}(Y^m) : \theta \in \Theta \subset \mathbf{R}^k\}$ とかけるk次元パラメトリックな確率モデルのクラスを考える。ここで、Θはk次元の実数値パラメータ空間である。データ系列Y^mが与えられたとして、最初に思い付くパラメータの推定量は**最尤推定量**である。これは、尤度$P_{\theta,k}(Y^m)$を最大にする推定量として定義される。$\hat{\theta}$をY^mからの最尤推定量として、$P_{\hat{\theta},k}(Y^m)$に対してY^mを符号化すれば確率的コンプレキシティが得られそうだと簡単に思えるかも知れない。ところが、$P_{\hat{\theta},k}$は確率分布をなさないという点に注意しなければならない。実際、各々の$P_{\hat{\theta},k}$はY^mに依存するので、$\sum_{Y^m \in \mathcal{A}^m} P_{\hat{\theta},k}(Y^m) > 1$となる場合が存在する。そこで、正規化して得られる$Y^m$の関数、$P_{\hat{\theta},k}(Y^m)/\sum_{Y^m} P_{\hat{\theta},k}(Y^m)$は$Y^m$に関する確率分布をなすことに注目すると、この分布に対するY^mの符号化が定義できる。これを**最尤符号化 (Maximum Likelihood Coding)**[4]·[11]とよぶ。さらには、その符号長で確率コンプレキシティを近似することが出来る。すなわち、

$$SC(Y^m : \mathcal{H}^{(k)}) \approx -\log\left(\frac{P_{\hat{\theta},k}(Y^m)}{\sum_{Y^m} P_{\hat{\theta},k}(Y^m)}\right). \quad (3)$$

(3)の右辺の量を$I(Y^m : \mathcal{H}^{(k)})$とかくことにする。この値に関しては次が知られている。

定理 1 *[11]* θのほとんどいたるところで最尤推定量に関して中心極限定理が成立するものと仮定する。このとき次式が漸近的に成り立つ。

$$I(Y^m : \mathcal{H}^{(k)}) \quad (4)$$
$$= -\log P_{\hat{\theta},k}(Y^m) + \frac{k}{2}\log\frac{m}{2\pi} + \log\int\sqrt{|I(\theta)|}d\theta + o(1).$$

ここに、$I(\theta) = E_{\theta,k}[\frac{\partial^2 \log P_{\theta,k}(Y)}{\partial\theta_i \partial\theta_j}]$は$\theta$における*Fisher*情報量行列とよばれる量であり*(*$E_{\theta,k}$は$P_{\theta,k}(Y^m)$に関する平均を表す*)*、$|I(\theta)|$はその行列式を表す。$o(1)$は$\lim_{m\to\infty} o(1) = 0$となる量である。

このように、最尤符号化の符号長が評価されたのだが、これによる確率コンプレキシティの近似(3)は果してタイトなものなのであろうか？ 結論から先に言えばYesである。実際に以下の不等式が成立する。

定理 2 *[9]*$L(Y^m)$を任意の語頭符号化の符号長関数とする。データY^mが確率分布$P_{\theta,k} \in \mathcal{H}^{(k)}$に従って発生するとき、定理*1*と同じ条件のもとで、すべての$\varepsilon > 0$に対し、漸近的に測度が0となる実数値パラメータの集合を除いた全てのθについて次式が成立する。

$$E_{\theta,k}[L(Y^m)] \geq H_m(P_{\theta,k}) + \left(\frac{k}{2} - \varepsilon\right)\log m. \quad (5)$$

ここで、$H_m(P_{\theta,k}) \stackrel{\text{def}}{=} E_{\theta,k}[-\log P_{\theta,k}(Y^m)]$は$P_{\theta,k}(Y^m)$に関するエントロピーを表す。

ここで、$H_m(P_{\theta,k}) \approx E_{\theta,k}[-\log P_{\hat{\theta},k}(Y^m)] + \frac{k\log e}{2}$であることが知られているから、(4)と(5)を比較すると、$I(Y^m : \mathcal{H}^{(k)})$は真の分布に関する平均の意味で(5)の下界を$o(\log m)$の誤差以内で達成していることがわかる。ここに、$\lim_{m-\infty} o(\log m)/\log m = 0$である。よって、真の分布が$\mathcal{H}^{(k)}$の中に含まれている状況では、(4)の右辺の値は確率的コンプレキシティを平均として$o(\log m)$以内の精度でタイトに近似しているといえる。さらに注目すべきことに、(4)の右辺の第3項に関しては、実は、ベイズリスクのミニマックス戦略の立場からその最適性が証明されている[2]。

以上、近似(3)がタイトであることを見た。以下、連続なパラメータのみで指定されたモデルクラスに対しては、(4)で計算できる$I(Y^m : \mathcal{H}^{(k)})$の値そのものを、$Y^m$の$\mathcal{H}^{(k)}$に対する確率的コンプレキシティの定義として議論を進めて行こう。

3.3 確率モデルの最適選択

これまではパラメータの次数kを固定してきたが、kに関して入れ子構造をもつクラスの系列：

$$\mathcal{H}^{(1)} \subset \mathcal{H}^{(2)} \subset \cdots \subset \mathcal{H}^{(k)} \subset \mathcal{H}^{(k+1)} \subset \cdots \mathcal{H}^{(s)}$$

を考えて、その和集合を$\mathcal{H} = \cup_k \mathcal{H}^{(k)}$としよう。$k$に関する事前分布を$\pi(k)$として、データ系列$Y^m$の$\mathcal{H}$に対する確率的コンプレキシティ$SC(Y^m : \mathcal{H})$を2段階符号化で近似すると次のようになる。

$$SC(Y^m : \mathcal{H}) \approx \min_k\{I(Y^m : \mathcal{H}^{(k)}) - \log\pi(k)\}.$$

もっとも、kに関して特に事前知識がなければ、$\pi(k)$を一様分布に設定することにより$-\log\pi(k)$の項を無視して、単純に$I(Y^m : \mathcal{H}^{(k)})$の$k$に関する最小化の問

題として捉えることが出来る。その際の上式の右辺の最小化は以下に帰着できる。

$$\min_k \left\{ -\log P_{\hat{\theta},k}(Y^m) + \frac{k}{2}\log\frac{m}{2\pi} + \log \int \sqrt{|I(\theta)|}d\theta \right\}. \tag{6}$$

$\hat{\theta}$はY^mからの最尤推定値である。kはモデルクラスの複雑さを表す一種の指標であるが、真の分布を記述する最小のkが未知であるとして、ここでMDL原理を適用すると、(6)の右辺の最小を達成するようなkが最適なモデルのパラメータ次元ということになる。これがパラメータ次数選択における**MDL基準**[8],[9]と呼ばれるものである。MDL基準はしばしばBayes理論の立場から解釈されているが、最尤符号化からの導出において事前分布などのBayes的な仮定を一切おいていないことに注意しよう。

今、(6)の最小値を達成するkを$\hat{k}$としよう。もし、真の分布P^*のパラメータの次数がk^*であるとしたら、(6)と定理2から、漸近的には$\hat{k}$はk^*に確率収束することを示すことができる。この性質を**一致性**とよぶ。確率モデル選択基準の"良さ"はいろいろな尺度で評価され得るが, MDL基準の良さの1つは一致性にある。

3.4 具体例その1: ヒストグラムの推定

定義域を$\mathcal{X}=[0,1]$として、$\mathcal{X}$上の$k+1$分割の**ヒストグラム密度**とは、$\mathcal{X}$を$k+1$等分して、それぞれのセルをC_i $(i=1,\cdots,k+1)$として、C_iに入ったデータXに対する確率密度を$(k+1)\theta_i$で定めるような確率密度関数である$(0\le\theta_i\le 1,\ \sum_{i=1}^{k+1}\theta_i=1)$。$k+1$分割のヒストグラム密度全体のクラスを$\mathcal{H}_{HIS}^{(k)}$とかく。データ系列$Y^m=Y_1\cdots Y_m$が与えられたとして、そのうち$i$番目のセルに入ったデータの数を$m_i$とするとき$(m=\sum_{i=1}^{k+1}m_i)$、尤度は

$$\prod_{i=1}^{k+1}((k+1)\theta_i)^{m_i}$$

と計算できるから、θ_iの最尤推定量は$\hat{\theta}_i=m_i/m$であることがわかる。ただし、$0\log 0=0$とする。また、このときのFisher情報行列の行列式は$|I(\theta)|=1/\prod_{i=1}^{k+1}\theta_i$と与えられるから、$\int\sqrt{|I(\theta)|}d\theta$は

$$\int\prod_{i=1}^{k+1}\theta_i^{-1/2}d\theta=\frac{\pi^{k/2}}{\Gamma(k/2)}$$

のように計算できる。ここに、Γはガンマ関数を表す。結局、(4)の$I(Y^m:\mathcal{H}^{(k)})$は以下のように計算できる。

$$-\sum_{i=1}^{k+1}m_i\log\frac{m_i}{m}-m\log(k+1)+\frac{k}{2}\log\frac{m}{2\pi}+\log\frac{\pi^{k/2}}{\Gamma(k/2)}.$$

ここで$o(1)$の項は無視した。よって、与えられたデータに対して最良のヒストグラムの分割数を推定するには上式の値を最小にするkを求めれば良い。(注:上式にて、データの符号長を(−log(密度関数の尤度))と形式的に計算しているが、これは負の値をとる。$\mathcal{X}$が連続の場合は、本来は$\mathcal{X}$を離散化して、その上の確率分布に対して符号長を計算しなければならないのだが、上記のような形式的展開も許される。)

3.5 具体例その2:確率的規則の一括学習

$\mathcal{X}\subset\mathbf{R}^n$, $\mathcal{Y}=\{0,1\}$とする。今、入力変数$X\in\mathcal{X}$と出力変数$Y\in\mathcal{Y}$の組$D=(X,Y)$が独立に未知の真の分布$P(X,Y)=Q(X)P^*(Y|X)$に従って生成されているとする。$\mathcal{H}$を条件付確率モデルのクラスとし、データ系列$D^m=D_1\cdots D_m$ $(D_i=(X_i,Y_i))$が一括与えられたとして、$\mathcal{H}$から$P^*(Y|X)$に出来るだけ近い確率モデルを1つ選び出す問題を**一括学習の問題**とよぶ。

いま、$\mathcal{X}$を有限個の排反する領域$\{C_i:i=1,\cdots,k\}$に分け、$\mathcal{X}=\cup_{i=1}^k C_i$ $(C_i\cap C_j=\phi, i\neq j)$とし、$X$が領域$C_i$に入ったら確率$p_i$で$Y=1$を、確率$1-p_i$で$Y=0$を出力するような確率モデルを考える。このような確率モデルを**有限分割型の確率規則**[12]とよぶ。k個の有限分割で指定される有限分割型の確率規則のクラスを$\mathcal{H}_{FP}^{(k)}$とかく。データ系列D^mが与えられたとき、$\mathcal{H}_{FP}^{(k)}$のモデルについて、Xがi番目の領域に属するデータの数をm_i、そのうち$Y=0,1$であるデータの数をm_{i1},m_{i0} $(m_{i1}+m_{i0}=m_i)$とすると、尤度は$\prod_{i=1}^k\theta_i^{m_{i1}}(1-\theta_i)^{m_{i0}}$とかけるので、$\theta_i$の最尤推定値は$\hat{\theta}_i=m_{i1}/m_i$と計算される。また、Fisher情報行列の行列式は$|I(\theta)|=1/\prod_{i=1}^k\theta_i(1-\theta_i)$で与えられるから、$\log\int\sqrt{|I(\theta)|}d\theta=\log(\sqrt{\pi}/\Gamma(1/2))^k=0$と計算される。よって、MDL基準では$\mathcal{H}_{FP}=\cup_k\mathcal{H}_{FP}^{(k)}$の中で

$$-\sum_{i=1}^{k}\sum_{j=0,1}m_{ij}\log\frac{m_{ij}}{m_i}+\frac{k}{2}\log\frac{m}{2\pi} \tag{7}$$

を最小にする$\hat{P}(Y|X)$が最適モデルとして選ばれる。

一括学習の性能の良さを測る学習基準として、**確率的PAC(Probably Approximately Correct)学習基準**[12]というものが知られている。これは、真の分布$P^*(Y|X)$とアルゴリズムが推定する分布$\hat{P}(Y|X)$の距離を$d(P^*,\hat{P})$として、$0<\varepsilon,\delta<1$が与えられたもとで、$1-\delta$以上の確率で$d(P^*,\hat{P})<\varepsilon$となるのに必要最小のデータ数で(これを**サンプルコンプレキシティ**とよぶ)アルゴリズムを評価するものである。

このサンプルコンプレキシティはアルゴリズムの一種の収束速度の指標であり、$1/\varepsilon, 1/\delta$のオーダーとして小さければ小さいほど良い。

(7) を最小にするモデルを出力するようなアルゴリズムのサンプルコンプレキシティは、もし、あるk^*に対して真の分布P^*が$\mathcal{H}^{(k^*)}$に含まれている場合には、

$$O\left(\frac{k^*}{\varepsilon}\log\frac{k^*}{\varepsilon}+\frac{\ell(P^*)}{\varepsilon}+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right) \tag{8}$$

で与えられることが知られている [12]。ここに$d(P^*,\hat{P})=\int dXQ(X)\sum_{Y=0,1}(P^*(Y|X)^{1/2}-\hat{P}(Y|X)^{1/2})^2$はHellingerの距離であり ($Q$は$\mathcal{X}$上の分布を表す)、$\ell(P^*)$は$P^*$の有限分割の記述に必要な符号長である。(8) の値は$1/\varepsilon, 1/\delta, k^*$の関数として最小であることも知られており、MDL基準の一括学習におけるサンプルコンプレキシティの意味での最適性を保証している。ただし、真の分布P^*がどの$\mathcal{H}^{(k)}$にも含まれない場合の最適性は必ずしも保証されない。

4. 逐次的符号化と確率的予測

4.1 混合符号化

3節ではデータ系列が一括与えられたもとで確率的コンプレキシティを近似する符号化を考えてきた。ところが、データ系列が1つ1つ逐次的に与えられて、オンラインで符号化しなければならない状況というのも考えられる。本節では、このような状況下での、与えられた確率モデルのクラスに対するデータ系列の符号化 (これを**逐次的符号化**とよぶ) による確率的コンプレキシティの近似について考えてゆこう。

先ず最初に**混合符号化 (Mixture Coding)** あるいは **Bayes 符号化** [1]-[7] とよばれる逐次的符号化を紹介しよう。これは、確率モデルのクラス$\mathcal{H}^{(k)}=\{P_{\theta,k}(Y):\theta\in\Theta\subset\mathbf{R}^k\}$が与えられたとして、このクラスの要素全ての重みつき平均として定義される確率分布に対して符号化する方法で、しかもその重みの値が逐次的に変化していくというものである。具体的には、t番目のデータY_tを確率分布

$$\bar{P}_t(Y)=\int w(\theta|Y^{t-1})P_{\theta,k}(Y)d\theta \tag{9}$$

に対して符号化する。ここで重み$w(\theta|Y^{t-1})$は過去のデータ系列$Y^{t-1}=Y_1\cdots Y_{t-1}$からBayesの事後確率

$$w(\theta|Y^{t-1})=\frac{\pi(\theta)P_{\theta,k}(Y^{t-1})}{\int\pi(\theta)P_{\theta,k}(Y^{t-1})d\theta}$$

として計算するものとする。ここで$\pi(\theta)$は予め与えられた事前分布である。データ系列が$Y^m=Y_1,Y_2,\cdots,Y_m$の順で与えられたとして、混合符号化した場合のY^mの符号長は

$$\sum_{t=1}^{m}(-\log\bar{P}_t(Y_t))=-\log\int\pi(\theta)P_{\theta,k}(Y^m)d\theta \tag{10}$$

と計算できる。もし、確率モデルのクラスを$\mathcal{H}^{(k)}$の代わりに$\mathcal{H}=\cup_k\mathcal{H}^{(k)}$とすれば、$Y^m$に対する$\mathcal{H}$の混合符号化も、もう一段高いパラメータ$k$に関する階層を考慮することにより同様に定義できる。

4.2 予測的符号化

もう1つの代表的な逐次的符号化方法として**予測的符号化 (Predictive Coding)**[9] を紹介しよう。これは、確率モデルのクラス$\mathcal{H}^{(k)}$が与えられたとして、逐次的にパラメータθを過去のデータから推定しながら符号化する方法である。具体的には次のような符号化を行う。t番目のデータY_tを確率分布

$$\hat{P}_t(Y)=P_{\hat{\theta}_{t-1},k}(Y) \tag{11}$$

に対して符号化する。ここで$\hat{\theta}_{t-1}$はパラメータθの過去のデータ系列$Y^{t-1}=Y_1\cdots Y_{t-1}$からの推定値 (例えば、最尤推定値) である。データ系列が$Y^m=Y_1\cdots Y_m$の順で与えられたとして、予測的符号化によって符号化した場合のY^mの$\mathcal{H}^{(k)}$に対する符号長は、

$$-\sum_{t=1}^{m}\log P_{\hat{\theta}_{t-1},k}(Y_t) \tag{12}$$

と計算される。この量はしばしば、(Y^mの$\mathcal{H}^{(k)}$に対する) **予測的確率的コンプレキシティ**[9] と呼ばれる。

混合符号化と予測的符号化といった逐次的符号化に要する総符号長 (10) と (12) は、いずれもY^mの$\mathcal{H}^{(k)}$に対する確率的コンプレキシティの近似と考えてよい。実際、予測符号化において$\hat{\theta}_{t-1}$をY^{t-1}からの最尤推定量とし、真の分布P^*が$\mathcal{H}^{(k)}$に属すると仮定して、$P^*=P_{\theta^*,k}$とすると、(12) のP^*に関する平均は漸近的に以下で与えられることが知られている [16]。

$$H_m(P_{\theta^*,k})+\frac{k}{2}\log m+o(\log m). \tag{13}$$

混合符号化についても、データが独立でパラメータ空間がcompact-supportedである場合には、真の分布$P^*=P_{\theta^*,k}$が$\mathcal{H}^{(k)}$に属するならば、(10) のP^*に関する平均は以下で与えられることが知られている [1]。

$$H_m(P_{\theta^*,k})+\frac{k}{2}\log\frac{m}{2\pi e}+\log\frac{\sqrt{|I(\theta^*)|}}{\pi(\theta^*)}+o(1). \tag{14}$$

ここに、θ^*は真の分布を指定するパラメータである。

(13),(14) と (4) を比べると、逐次的符号化である予測的符号化、混合符号化に要する総符号長は、非逐次的符号化である最尤符号化に要するそれ (すなわち、確率的コンプレキシティそのものの値) と、平均として $o(\log m)$ 以内で漸近的に一致しているという興味深い事実が浮かび上がる。さらに、混合符号化でパラメータの事前分布を Jeffereys の事前分布 $\pi(\theta) = \frac{\sqrt{|I(\theta)|}}{\int \sqrt{|I(\theta)|}d\theta}$ に設定すると ($I(\theta)$ は Fisher の情報行列)、混合符号化の平均総符号長は最尤符号化のそれと $O(1)$ の項まで一致することがわかる。

4.3 逐次的確率的予測モデル

確率分布と符号化は表裏一体の関係であることから、非逐次的符号化が符号化自体の意味を離れて、確率モデルの最適選択や確率的規則の一括学習という問題に有効な戦略を与えていた。これと同様に、逐次的符号化は符号化の意味を離れても、**逐次的確率的予測**という学習問題 [5]·[16] に有効な戦略を提示する。

逐次的確率的予測の問題とは以下のような問題である。データが $Y_1, Y_2, \cdots$ と順番に与えられる状況のもとで、$t-1$ 番目のデータまでの系列 $Y^{t-1} = Y_1 \cdots Y_{t-1}$ が与えられた時点で、Y_t をもらう前に Y_t が従う確率分布を予測したいものとする。予測アルゴリズムは、与えられた確率モデルのクラス $\mathcal{H}$ を用いて、Y^{t-1} の関数として Y_t の分布を予測し (予測分布を $P_t(Y)$ とかく)、予測後正しい値 Y_t を教えてもらう。このとき予測誤差を**対数誤差**

$$-\log P_t(Y_t)$$

で測るものとする。これは実際に生起した Y_t に対して小さい確率を割り当てるような分布を予測した場合には大きな値をとるような損失関数である。このプロセスを t に関して逐次的に繰り返す。我々は任意のデータ系列 Y^m に対してその**累積予測誤差**

$$\sum_{t=1}^{m}(-\log P_t(Y_t))$$

が出来るだけ小さくなるような予測アルゴリズムを設計したい。ここで、各時刻での予測誤差は予測分布 $P_t(Y)$ に対する Y_t の符号長という解釈が出来るので、累積予測誤差最小の問題はまさしく逐次的符号化による符号長最小化の問題に他ならない。

そこで、確率モデルのクラス $\mathcal{H}^{(k)} = \{P_{\theta,k}(Y) : \theta \in \Theta \subset \mathbf{R}^k\}$ が与えられたとして、混合符号化に対応して、各時刻 t で (9) の分布を出力する予測アルゴリズム (これを**Bayes 予測アルゴリズム**とよぶ) が考えられる。また、予測的符号化に対応して、各時刻 t で (11) の分布を出力する予測アルゴリズム (これを**最尤予測アルゴリズム**とよぶ) も考えられる。

もし、真の分布 P^* が $\mathcal{H}^{(k)}$ に属するならば、(13) と (14) からわかるように、上の 2 つの予測アルゴリズムに対する平均累積予測誤差はいずれも $H_m(P_{\theta^*,k}) + \frac{k}{2}\log m$ に $o(\log m)$ 以内の誤差で収まり、平均予測誤差の下界式 (5) と $o(\log m)$ の誤差範囲内で漸近的に一致することがわかる。この意味で上の 2 つのアルゴリズムは最適であるといえる。このように、逐次的符号化を用いた確率的コンプレキシティの近似を考えることによって、最適な逐次的確率予測アルゴリズムが具体的に設計できるのである。

5. 確率的コンプレキシティの発展

以上見たように、確率的コンプレキシティは確率モデル選択や逐次的確率予測の問題に対して有効な戦略設計指針を与えている。しかしながら、現実的な学習問題への適用において幾つか問題が残っている。

1. 一般化確率的コンプレキシティ

問題の 1 つは、統計的決定理論の立場から見ると、確率的コンプレキシティの定義において、モデルのクラスは確率モデルのクラスであり、損失関数は対数誤差を用いるといった制限が与えられていたことである。ところが、現実的な学習の問題を考えると、モデルのクラスが一般の実数値関数であり、損失関数も自乗誤差や絶対誤差などを含む一般的な損失関数を考えなければならない状況が多い。そのような状況に対応して確率的コンプレキシティの概念を一般化して、より広範囲の学習問題に適用できるようなアルゴリズムの設計と解析の理論を推し進めた研究も最近発展している [13]·[17]。

2. ランダム化による近似

問題のもう 1 つは、確率的コンプレキシティの近似過程において、しばしば計算論的困難が伴うということである。例えば、確率的モデルのクラスが有限であるが指数的多数である場合や、隠れ変数を伴うモデルなどの複雑な連続パラメータ構造をもつ場合は、混合分布 (9) を求める際、あるいは最尤符号化 (3) で最尤推定値を求める際に計算量が指数的になったり、あるいは解析的には計算不可能という事態に陥る。その場合に計算的困難を克服していく有望なアプローチの 1

	符号化の方法		統計推論/学習の問題	有効性の理由
確率的コンプレキシティ	非逐次的符号化	2段階符号化	確率モデルの最適選択	一致性、サンプルコンプレキシティの最小性
		最尤符号化		
	逐次的符号化	混合符号化	逐次的確率的予測	累積予測誤差の最小性
		予測符号化		

表 1: 符号化と統計推論／学習問題

つはマルコフチェイン モンテカルロ法などのランダム化手法を用いることである。この手法に基づいて、計算量を考慮した具体的な確率的コンプレキシティの近似理論の研究が進んでいる [15]-[18]。

6. おわりに

本稿では、確率モデルのクラスに対するデータの最小符号長として確率的コンプレキシティを導入し、それを最良近似するための符号化の過程から、統計的推論や学習に有効な戦略が生まれることを見てきた。確率的コンプレキシティを近似するための符号化方法として、逐次的符号化と非逐次的符号化の2種類があることを示した。非逐次的符号化としては2段階符号化、最尤符号化などがあり、符号化自体の意味を離れても、それらは確率モデルの最適選択や一括学習方式のアルゴリズムの設計指針を与えている。それらの有効性はパラメータ次数推定の一致性や一括学習のサンプルコンプレキシティの最小性によって保証されていることを見た。また、逐次的符号化としては混合符号化、予測的符号化などがあり、符号化自体の意味を離れても、それらは逐次的確率予測アルゴリズムの設計指針を具体的に与えている。それらは対数誤差で測った累積予測誤差を最小にするという意味で最適であることを見た。以上をまとめたのが表1である。

本稿で扱わなかった重要な統計的推論の問題の1つに**仮説検定**と呼ばれるものがある。確率的コンプレキシティはユニバーサルな仮説検定問題に対してもやはり有効な検定方式を提示する。この方式の理論的妥当性は**PAD 学習 (Probably Almost Discriminative Learning)** という文脈の中で証明されている [14]。

参考文献

[1] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory* **IT-36** (1990), 453-471.

[2] B.S. Clarke and A.R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," to appear in *JSPI.*

[3] T.M. Cover and J.A. Thomas, "Elements of Information Theory," Wiley-Interscience, 1991.

[4] L.D. Davisson, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory* **IT-29**, 2 (1983), 211-215.

[5] A. Dawid, "Statistical theory: the prequential approach," *J. R. Stat. Soc. A* (1984), 278-292.

[6] D. Haussler, "Generalizing the PAC model for neural net and other learning applications," *Inform. Comput.*, 100 (1992), 78-150.

[7] T. Matsushima, H. Inazumi, and S. Hirasawa, "A class of distortionless codes designed by Bayes decision theory," *IEEE Trans. Inform. Theory,* **IT-37**, 5 (1991), 1288-1293.

[8] J. Rissanen, "Modeling by shortest data description," *Automatica*, **14** (1978), 465-471.

[9] J. Rissanen, *Stochastic Complexity in Statistical Inquiry,* World Scientific, 1989.

[10] J. Rissanen and B. Yu, "MDL learning," in *Progress in Automations and Information Systems,* Springer Verlag, 1992.

[11] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory,* **IT-42**, 1 (1996), 40-47.

[12] K. Yamanishi, "A learning criterion for stochastic rules," *Machine Learning* **9** (1992), 165-203.

[13] K. Yamanishi, "Generalized stochastic complexity and its applications to learning," in *Proceedings of the 1994 Conference on Information Science and Systems,* (1994), vol.2, pp.763-768.

[14] K. Yamanishi, "Probably almost discriminative learning," *Machine Learning,* **18** (1995), 23-50.

[15] K. Yamanishi, "Randomized approximate aggregating strategies and their applications to prediction and discrimination," in *Proc. of COLT95*, (1995), pp.83-90.

[16] K. Yamanishi, "A loss bound model for on-line stochastic prediction algorithms," *Inform. Comput.*, **119**, 1, (1995), 39-54.

[17] K. Yamanishi, "On-line maximum likelihood prediction with respect to general loss functions," to appear

in *Jr. Comput. Sys. Sci.*, (1995). An extended abstract appeared in *Proc. of EuroCOLT'95*, Springer, (1995), pp.84-98.

[18] K. Yamanishi, "A randomized approximation of the MDL for stochastic models with hidden variables," to appear in *Proc. COLT'96*, (1996).

[19] 山西、韓、"ＭＤＬ入門：情報理論の立場から，"人工知能学会誌　vol 7, No 3, May (1992), 45-52.