情報論的学習理論の現状と展望

山西 健司
NEC情報通信メディア研究本部
yamanisi@ccm.cl.nec.co.jp

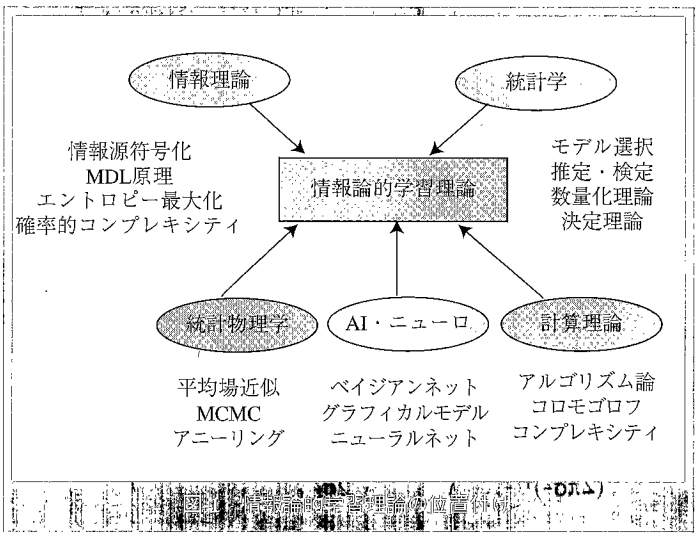
■情報論的学習理論とは■

はじめに断っておくが，「情報論的学習理論」という学問分野が確立しているわけではない．これは機械学習という問題を通じて，情報理論，統計学，統計物理学，計算機科学，応用分野が垣根を越えて自由に交流するための場につけられた名称に過ぎない．

そもそも，機械学習とはデータに潜む有用な知識を効率的に自動発見するための技術である．データの大規模性に対するScalabilityが強調されるとデータマイニングと呼ばれている．一方で，情報理論や統計学などでは，古くからデータから因果関係を抽出したり，データに含まれる情報の量を測るための数理的手法を蓄積してきた．たとえば統計的モデル選択やユニバーサル符号化では，未知の情報源の出力系列から情報源モデルを推測したり，また，それを基に効率よい符号化を行うための方法論が展開されている．しかし，だからといって機械学習のすべての問題が古典的な情報理論や統計学の枠組みの中で解決できるとは限らない．事実，機械学習で用いられる複雑な知識表現系（たとえば，ベイジアンネット，ニューラルネット，Support Vector Machine等）やこれらを効率的に学習するためのアルゴリズム（たとえば，EMアルゴリズム，アニーリング，ブースティング等）は統計物理学や計算機科学と結びついて，逆に情報理論や統計学に新しい刺激を与えている．また，こうした理論は画像処理，言語処理，生体情報処理，信号処理，パターン認識等にインパクトを与えている．

つまり，機械学習をめぐって，さまざまな分野が独自の強みに訴えながらいろいろな角度から論じているのである．しかし，いずれも学習を「情報を最大限に抽出するプロセス」としての情報論的視点で捉え，数理的に分析しているという共通項を見出すことができるであろう．そこで，上記分野の相互交流を通じて学習問題の本質が一層の深みをもって見えてきそうである．しかしながら，これまで正面きってそれぞれの考え方を真に有機的に交流させ相互理解しようとする試みは必ずしも積極的に行われてきたとはいえないのではないかと思われる．「情報論的学習理論」とは，情報論的手法（情報を数理的に扱う手法という意味）を軸にしてそういった交流を促進するためのボーダレスな統計的知識情報処理の枠組みとでもいおうか（図-1）．

1998年にIBIS（Information-Based Induction Sciences: 邦題「情報論的学習理論ワークショップ」）という会議が，実験的交流の機会として初めて国内で開催された（IBISは英語で鳥のトキを意味する）．従来の計算論的学習理論や機械学習の会議よりも情報理論，統計学，統計物理学のコアの部分にどっぷりと踏み込みながら学際的



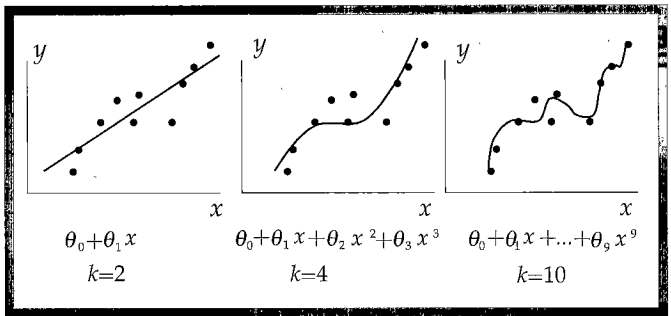


図-2 回帰モデルの次数推定

な盛り上がりを見せたのである．その後，IBISは規模を拡大しながら毎年開催されるようになり，各学会論文誌等でも特集記事が組まれるようになった[10)～15)]．機械学習技術がデータマイニングなどのビジネス応用に直結する昨今において，その基礎にあたる理論の本質を理解したい，特に底流としての情報理論や統計学とのかかわりをしっかりと押さえておきたいという人々が増えるのも当然の成り行きである．情報論的学習理論の最近の発展はそんな事情を反映しているのかもしれない．

本稿では，過去3年間のIBISでの発表の主なトピックスを中心に情報論的学習理論の現状を概観する．トピックスとしては以下のようなものが挙げられる．

- 一括学習と統計的モデル選択
- 逐次的予測問題と確率的コンプレキシティ
- 最適化アルゴリズムと実践ベイズ法
- アンサンブル学習
- その他学習理論（非正則系，量子論的能動学習等）
- 応用分野

■統計的モデル選択■

まずデータが一括して与えられたもとで統計的な知識を学習する問題（**一括学習**）を考える．データの生成機構を表す統計的モデルが複数あったとして，その中から与えられたデータ列の生成機構を説明するものとして最も適切なものを選択する問題を**統計的モデル選択**の問題と呼ぶ．これは最も基本的な機械学習問題の1つである．

今，説明変数をx，目的変数をyとして，xが与えられときのyの条件付確率分布の集合として，ある構造を表す離散パラメータM（以下，モデルと呼ぶ）で指定されたパラメトリックな集合を$\mathcal{C}_M=\{p(y|x:\theta,M):\theta\in\Theta_M\}$と書く．ここに$\Theta_M$はパラメータ空間，$p(y|x:\theta,M)$は連続値パラメータ$\theta$で指定された条件付確率分布を表す．たとえば$x,y$は共に1次元とし，$k$次の回帰モデル:

$$p(y|x:\theta,k)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{(y-(\theta_0+\theta_1 x+\cdots\theta_{k-1}x^{k-1}))^2}{2\sigma^2}\right)$$

を考えると（σは正の定数），モデルMはパラメータの次元kであり，$\theta=(\theta_0,\cdots,\theta_{k-1})$である．

$D=(x,y)$の実現値として入出力対の事例データ列$(x_1,y_1),\cdots,(x_m,y_m)$を観測したとき，これに基づいて$\mathcal{C}_M(M=1,2,\ldots)$の中から最適なモデルを選択したい．上の例では，統計的モデル選択とは最適な次元$M=k$を求めることである（図-2）．

このために，まず期待平均対数尤度の最大化を目的とした統計的モデル選択規準を考えよう．以下では，表記の簡略化のために，変数xを省略し，yのみを残す．$M=k$の場合を考えて，$\mathcal{C}_k=\{p(y|\theta,k):\theta\in\Theta_k\}$と書く．データの生成は独立であると仮定する．また，長さmの事例データ列を$y^m=y_1\cdots y_m$と書き，$\hat{\theta}$をy^mからのθの最尤推定量（$\prod_{t=1}^m p(y_t|\theta,k)$を最大にする$\theta$）とする．今，**期待平均対数尤度**の$(-m)$倍を

$$mE_{y^m}E_{y'}[-\log p(y'|\hat{\theta},k)] \qquad (1)$$

のように定義する．ここでE_{y^m}は事例データ列y^mの生起に関する期待値を，$E_{y'}$は未来のデータy'の生起に関する期待値を表す．(1) は$p(y'|\hat{\theta},k)$で$\mathcal{C}_k$を代表させ，その負の対数尤度で未来のデータに関する損失（これを**対数損失**と呼ぶ）で測った場合の**期待平均対数損失**である．目的は期待平均対数損失を最小化することである．

(1) はパラメータの次元kに関する関数であるから，これを最小化するようなkを求めればよい．しかし，式 (1) は，“真の分布”（データが従う分布）に関する期待値操作を含んでおり，真の分布は未知であるから実際には計算できない．そこで，真の分布が$\mathcal{C}_k$に含まれると仮定し，(1) を事例データからの漸近不偏推定量：

$$-\sum_{t=1}^{m}\log p(y_t|\hat{\theta},k)+k \qquad (2)$$

で置き換え，この（kに関する）最小化という形で実現した統計的モデル選択規準を**AIC**（Akaike Information Criterion）[1)] と呼ぶ．AICはモデルのデータに対する適合度を表す第1項とモデル自体の複雑さを表す第2項をバランスした上でkを最適化する戦略を与える．

一方，情報理論の立場からは，「モデル自体の符号化も含めて最も短くデータを符号化できるようなモデルが最良である」といった統計的モデル選択規準が提案されている．これは**MDL規準**（Minimum Description Length Criterion）と呼ばれている．具体的には，データ列y^mを$\mathcal{C}_k$に対する最尤符号化と呼ばれる情報源符号化を行った場合の符号長

$$-\sum_{t=1}^{m}\log p(y_t|\hat{\theta},k)+\frac{k}{2}\log\frac{m}{2\pi}+\log\int\sqrt{|I(\theta)|}d\theta \qquad (3)$$

をkに関して最小化するという形で実現できる．ここに$I(\theta)=E_\theta[-\partial^2\log p(y|\theta,k)/\partial\theta_i\partial\theta_j]$はモデルの構造を反映するFisher情報行列と呼ばれるものである．(3) の量はy^mの$\mathcal{C}_k$に関する**確率的コンプレキシティ**（Stochastic

Complexity）と呼ばれている．この量はモデルクラス$\mathcal{C}_k$を通じてデータy^mを情報理論的な要請（一意的に復号化できるという要請）のもとで符号化した際に，最も短く符号化できる場合の符号長という意味を持つ．

上述のAICやMDLは，確率的決定リストや決定木等の確率的分類ルールの学習，ニューラルネットのノード数の最適化，ベイジアンネットの構造学習，ヒストグラムのセル数の最適化等に幅広く応用され，与えられた表現系の中で適切な自由度や複雑さを決定する際に有効な戦略を与えている．たとえば，図-3では正例○と負例×を分類する分割ルールをいくつか表しているが，これらの中から最適なルールを見つける場合などにも有効である．

AICとMDLの形式的な違いは，モデルの複雑さを表すペナルティ項がAICではkであるのに対し，MDLでは$(k/2)\log(m/2\pi)+\log\int\sqrt{|I(\theta)|}d\theta$となっていることである．果たしてそれらはどう使い分ければよいのだろうか？　これを期待平均対数損失，一致性，累積予測損失の観点からまとめよう．

1）期待平均対数損失：AICはもともと期待平均対数損失を最小化する目的で導入されたものであるから，予測や制御のように，推定値が得られるたびに1つ先のデータに対して分布予測しようというときにはAICが適切であると思われる．しかしここで注意したいのは，期待平均対数損失（1）の不偏推定量であるAICを最小化するkを選んだところで，それは真の期待平均対数損失を最小化する理論的保証はないということである．

どの基準が期待平均対数損失最小化の意味で最良であるかは，モデルと真の分布との関係で決まるのであり，どの基準が一様に最良であるとはいえない．モデルクラスとして自己回帰モデルを選んだとき，真の分布が無限個のパラメータを持った極限として表される場合には，AICが期待平均対数損失を最小化する唯一の基準であることが知られている[5]．一方，モデルクラスとして$\exp(\sum_{i=1}^{k}\theta_i\phi_i(y))/c(\theta)$（$\phi_i(y)$は直交基底，$\theta$は楕円上束縛される）の形のクラスを選んだとき，真の分布に関するある正則条件のもとでMDLが最良であることが知られている．

2）一致性：真の分布が有限のパラメータ次元k^*を持つとして，サンプル数を無限大にしたときにk^*を選択しない確率をゼロにできる性質を**一致性**と呼ぶ．自己回帰モデルやその他のモデルに対して，一致性を保証するためのモデル選択基準のペナルティ項は$\Omega(\log\log m)$であることが知られている．したがって，AICでは極限においてもk^*より大きい次元を推定する確率はゼロにはならない．一方，MDLは一致性を持つので，次元そのものの推定に興味がある場合はMDLを用いればよい．しかし，一致性を要求するだけではMDLである必要はなく，単に$\log\log m$をペナルティ項とすればよい．その意味で一致性はMDLを支持する決定的なものではない．

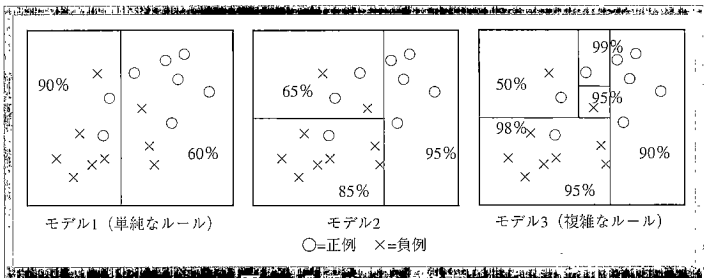


図-3　分類ルールの学習

3）累積予測損失：モデル選択を離れて逐次的予測問題を考えた場合，確率的コンプレキシティはミニマックスリグレットと呼ばれる基準のもとで最適な予測戦略の累積予測損失の値としての意味を持つ（次章参照）．

モデル選択規準を使い分けるには，以上のように規準を評価するメタな規準を明確に定めなければならない．

■逐次的予測問題■

本章では，データ列を$y^{t-1}=y_1,\ldots,y_{t-1}$まで観測して，次のデータ$y_t$を観測する前に，その生起確率分布を何らかの方法で$\hat{p}(\cdot|y^{t-1})$として予測するような場合を考えよう（たとえば，値域が{晴れ，雨}であり，過去の天気から明日の晴れの確率を予測するような場合）．予測の後には実現値y_tを受け取って，予測損失を**対数損失**$-\log\hat{p}(y_t|y^{t-1})$で測る．たとえば，晴れの確率を0.8として予測したとすると，実際に晴れた場合の予測損失は$-\log 0.8=0.223$，雨が降った場合には$-\log 0.2=1.609$と計算する．このような過程が時刻とともに繰り返されるとする．

このように各時刻tで予測確率分布を出力するアルゴリズム$\mathcal{A}$を**逐次的予測アルゴリズム**と呼ぶ．全データ数をmとしたとき，**累積予測損失**$\sum_{t=1}^{m}(-\log\hat{p}(y_t|y^{t-1}))$を最小にしたい．このために，どのように$\mathcal{A}$を設計すればよいであろうか？　累積予測損失はまた，逐次的に符号化を行ったときの総符号長（これを予測符号長と呼ぶ）と見なすことができるので，この問題は予測符号長を最短にする逐次的な符号化の設計問題と同一視できる．

累積損失を絶対的に最小化するアルゴリズムを構成することは不可能であるので，k次元パラメータで指定された確率分布の集合$\mathcal{C}_k=\{p(\cdot|\theta,k)\}$を導入して，これに対して相対的に評価する．そこで逐次的予測アルゴリズム$\mathcal{A}$の$\mathcal{C}_k$に対する**相対累積予測損失**を次式で定める．

$$R(\mathcal{A}, y^m : \mathcal{C}_k)$$
$$= \sum_{t=1}^{m} (-\log \hat{p}(y_t | y^{t-1})) - \inf_{\theta} \sum_{t=1}^{m} (-\log p(y_t | \theta, k))$$

これは$\mathcal{C}_k$上の最良の1点による割り当てに対する$\mathcal{A}$の累積予測損失の相対値である．天候予測の場合は，$\mathcal{C}_k$として$p(y=\text{晴れ})=\theta$ $(0 \le \theta \le 1)$といった1次元のモデルを考えると，右辺の第2項は$\hat{\theta}\,(=m_1/m : m_1$は晴れが起こった日数)に関する対数損失の値になる．このような最良の1点$\hat{\theta}$は全データが得られなければ求められないので逐次的予測アルゴリズムでは達成できないことに注意する．さらに**ミニマックスリグレット**を

$$R_m(\mathcal{C}_k) = \inf_{\mathcal{A}} \sup_{y^m} R(\mathcal{A}, y^m : \mathcal{C}_k) \tag{4}$$

で定める．これは相対累積予測損失のデータ系列に関する最悪値を考え，これがアルゴリズムのとり方によってどこまで小さくできるかを評価した量である．

そこでミニマックスリグレットを達成するアルゴリズムは何で，その値はいくらか？　が問題になる．このようにデータの発生源に関してまったく仮定を置かず，ワーストケースで評価する手法は現在の学習理論の主流となっており，平均的な場合を解析することが主流であった情報理論と統計学とは異なる新しい流れである．

式 (4) の評価基準のもとで有効なアルゴリズムの1つとして**ベイズアルゴリズム**と呼ばれるものがある．これは時刻tにおける予測分布を

$$\hat{p}(\cdot | y^{t-1}) = \int p(\cdot | \theta, k) p(\theta | y^{t-1}) d\theta$$

と計算する．ここに，$p(\theta | y^{t-1}) = p(\theta) \prod_{j=1}^{t-1} p(y_j | \theta, k) / \int p(\theta) \prod_{j=1}^{t-1} p(y_j | \theta, k) d\theta$ はθの事後確率であり，$p(\theta)$はθの事前分布である．天候予測の場合は，$p(\theta)$を一様分布とおくと簡単な計算からt日目が晴れの予測確率は$t_1/(t+1)$ $(t_1-1$は過去$t-1$日に晴れだった日数)で与えられることが分かる．ベイズアルゴリズムに対して，累積予測損失は以下のように計算できる．

$$\sum_{t=1}^{m} (-\log \hat{p}(y_t | y^{t-1})) = -\log \int p(\theta) \prod_{t=1}^{m} p(y_t | \theta, k) d\theta \tag{5}$$

実は，$p(\theta)$を適当におくことによりミニマックスリグレットを達成し，その値，つまり式 (5) の右辺は確率的コンプレキシティ (3) に一致するのである．このとき，

$$R_m(\mathcal{C}_k) = \frac{k}{2} \log \frac{m}{2\pi} + \log \int \sqrt{|I(\theta)|}\, d\theta + o(1)$$

となることが知られている．すなわち，確率的コンプレキシティ (以下，SCと略記) はミニマックスリグレットを達成する累積予測損失の値として特徴付けられる．よって，SCはモデル選択と逐次的予測の両面から最適学習戦略の設計に本質的にかかわる統計量であるといえる．

上記の逐次的予測問題はA) 確率分布の予測について，しかもB) 予測損失を対数損失で測る場合のものであった．この枠組みを

A)' 実数値関数による決定的な予測について，

B)' 一般の損失関数を用いて予測損失を測る

といった一般の統計的決定理論の枠組みへ拡張することができる．

今，xを説明変数，yを目的変数として，$\mathcal{F}_k = \{f_\theta(x) : \theta \in \Theta_k\}$を$k$次元パラメトリックな実数値関数のクラス，$L$を一般の損失関数 (たとえば，2乗損失は $L(y, \hat{y}) = (y-\hat{y})^2$で与えられる) とする．このとき，**拡張型確率的コンプレキシティ** (Extended Stochastic Complexity: 以下，ESCと略記) を次式で定義する[7]．

$$-\frac{1}{\lambda} \log \int p(\theta) \exp\left(-\lambda \sum_{t=1}^{m} L(y_t, f_\theta(x_t))\right) d\theta \tag{6}$$

$$\approx \min_{\theta} \sum_{t=1}^{m} L(y_t, f_\theta(x_t)) + \frac{k}{2\lambda} \log m + O(1) \tag{7}$$

ここにλは損失関数Lによって定まる定数である．対数損失の場合は$\lambda=1$であり，(6) は (5) に一致する．

上述の一般的な枠組みで逐次的予測問題を考え，ミニマックスリグレット (4) に相当する量を考えると，これを達成する予測アルゴリズムの累積予測損失はESCに漸近的に一致することが証明されている[7]．そのときのアルゴリズムはベイズアルゴリズムの一般化としての**Aggregating**アルゴリズムとして導かれる[6]．一方，一括学習問題としての一般的な損失関数に関する関数推定問題を考えると，ESCから (7) を最小化するようなkを選ぶような規準が得られる (これを**MLC** (Minimum L Complexity) アルゴリズムと呼ぶ)．そのような規準はMDLの一般化であり，実際，期待平均対数損失を一般化した統計的リスクという尺度に関して，これまで知られている関数推定の中で最小値を達成することも知られている[7]．よって，SCと同様にESCは推定と予測の両面から最適戦略を組み立てるための統一的な学習アルゴリズムの設計指標となっているのである (図-4)．

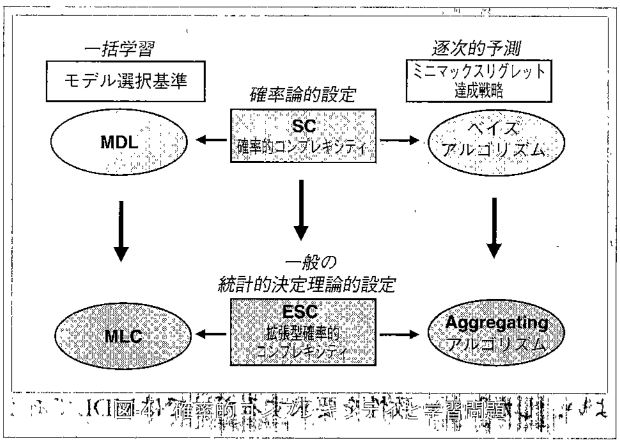


図-4 確率的[illegible]と学習問題

■最適化アルゴリズム■

機械学習の問題の多くは，尤度や事後確率や上述の情報量規準などを目的関数として，最大値あるいは最小値を解くことに帰着される．その際，大自由度かつ大量データに対して効率的に計算を行うための最適化アルゴリズムの設計が重要となる．

尤度を目的関数として，これを最大化するパラメータ推定問題（**最尤推定問題**）を考えよう．例として，今，データyがk個の異なる確率密度関数の**混合モデル**

$$p(y|\theta)=\sum_{i=1}^{k}\theta_i p_i(y) \tag{8}$$

に従って発生しているとする．$\theta_i>0$ $(i=1,\cdots, k)$かつ$\sum_{i=1}^{k}\theta_i=1$である．$\theta=(\theta_1,\cdots,\theta_k)$とおく．観測データ列$y^m=y_1\cdots y_m$を得たとき，最尤法によってパラメータ$\theta$を推定することを考えよう．対数尤度を

$$L(\theta|y^m)=\log\prod_{t=1}^{m}\sum_{i=1}^{k}\theta_i p_i(y_t)$$

と定める．そこで，最尤推定を行うには尤度方程式

$$\partial L(\theta|y^m)/\partial\theta=0$$

を解かなければならない．しかし，この場合，多変数の非線型方程式となるので，解析解を得るのは難しい．そこで，これを効率的に解くための方法が必要となる．中でも最も基本的なのが**EMアルゴリズム**である[2]．

EMアルゴリズムの鍵となるのは**隠れ変数**（直接観測できない変数）を導入することである．上の混合モデルの場合，観測データyが実際には混合要素の1つから生じたものと見なし，どの分布要素から生じたかを表すインデックスxを隠れ変数とする．xとyの同時確率は$\forall i$ $p(x=i, y|\theta)=\theta_i p_i(y)$のように単純な掛け算で表される．そこで，この同時確率に対する対数尤度関数の，隠れ変数に関する条件付期待値を

$$Q(\theta|\theta')=\sum_{x^m}p(x^m|y^m,\theta')\log\prod_{t=1}^{m}p(x_t,y_t|\theta)$$
$$=\sum_{t=1}^{m}p(x_t|y_t,\theta')\log p(x_t,y_t|\theta)$$

として定める．ここに，事後確率$p(x_t|y_t,\theta')$は

$$p(x_t|y_t,\theta')=\frac{p(x_t,y_t|\theta')}{\sum_{i=1}^{k}\theta_i' p_i(y_t)} \tag{9}$$

で与えられる．EMアルゴリズムは，$Q(\theta|\theta')$の計算（Expectation）とそのθに関する最大化（Maximization）を反復するプロセスとして与えられる．

EMアルゴリズム：

初期値$\theta^{(0)}$を与えて以下を収束するまで反復する．

E-Step：$Q(\theta|\theta^{(j)})$を計算する．

M-Step：$\theta^{(j+1)}=\arg\max_\theta Q(\theta|\theta^{(j)})$

$j\leftarrow j+1$.

ここで，反復数jが増大するにつれ$L(\theta^{(j)}|y_m)$は単調増大し，やがて極大値に収束することが知られている．

EMアルゴリズムのポイントは条件付期待対数尤度の計算が，隠れ変数を導入することによりたちまち容易になるといった構造にある．このような構造を持つモデルとしては，他に**因子分析モデル**や，**隠れマルコフモデル**，時系列解析で登場する**状態空間モデル**等がある．

EMアルゴリズムの問題点は大域的最適性を持たないことである．そこで，統計物理学にヒントを得て，条件付期待対数尤度の計算に温度を導入して大域的最適性を実現した**シミュレーティッドアニーリング**や**確定的アニーリング**が登場している．情報論的学習理論ではEMアルゴリズムの改良やデータマイニングやテキストマイニングにおける応用等が幅広く展開されている．

最尤推定とともに学習計算に頻繁に登場するのがベイズの事後確率である．データについて独立なモデル$p(y|\theta)$を用いた場合，データ列y^mに対するθの事後確率密度は

$$p(\theta|y^m)=\frac{p(\theta)\prod_{t=1}^{m}p(y_t|\theta)}{p(y^m)}$$

で与えられる．ここに$p(\theta)$はθの事前確率密度であり，

$$p(y^m)=\int p(\theta)\prod_{t=1}^{m}p(y_t|\theta)d\theta \tag{10}$$

である．$p(\theta|y^m)$が求められれば，これからのランダムサンプリングによってθの推定値が求められる．また，前章で述べたベイズアルゴリズムの計算を実行することができる．

ここで問題なのは，パラメータ数やデータ数が多くなった場合，(10)の計算が困難になり，事後確率密度が実際上計算できなくなるという点である．しかし最近では，この問題に対して統計物理学や計算統計学に基づいて，**マルコフ連鎖モンテカルロ法（MCMC），平均場近似，Variational Bayes法，Belief Propagation**等の高精度高速近似計算法が生まれている．これらは誤り訂正符号の復号化や画像復元等に応用されている[14]．

最も基本的なのがMCMCである．これを混合モデル(8)を例に説明しよう．今，事前分布$p(\theta)$を一様分布とすると，周辺分布(10)の計算は困難である．しかしながら前述の隠れ変数xを導入すると条件付確率$p(x^m|y^m,\theta)$や$p(\theta|x^m,y^m)$は次のように容易に計算できる．

$$p(x^m|y^m,\theta)=\prod_{t=1}^{m}p(x_t|y_t,\theta) \tag{11}$$

$$p(\theta|x^m,y^m)=D(m_1+1,\ldots,m_k+1) \tag{12}$$

ここに，(11)の$p(x_t|y_t,\theta)$は(9)で計算される．$D(m_1+1,\cdots,m_k+1)$は$\theta_1^{m_1}\cdots\theta_k^{m_k}$に比例する密度を持った

Dirichlet分布と呼ばれるものである．m_iはx^mの中で$x=i$が現れた数である．MCMCは，隠れ変数とパラメータをそれぞれ(11)と(12)に従って交互にランダムサンプリングを繰り返すプロセスとして与えられる．

マルコフ連鎖モンテカルロ法（MCMC）：
初期値$\theta^{(0)}$を与えて以下を反復する．
$\tilde{x}^m$を$p(x^m|y^m,\theta^{(j)})$に従ってサンプリングする．
$\theta^{(j+1)}$を$p(\theta|\tilde{x}^m, y^m)$に従ってサンプリングする．
$j \leftarrow j+1$.

jの増大につれ，$\Pi^{(j)}$の生起確率は事後確率$p(\theta|y^m)$に収束していくことが知られている．これによって事後確率分布を近似的に実現できるのである．

MCMCのポイントは，事後確率$p(\theta|y^m)$が容易に計算できなくても，隠れ変数xを導入することにより，条件付確率$p(x^m|y^m,\theta)$, $p(\theta|x^m, y^m)$が容易に計算できるようになるモデルの構造にある．

Variational Bayes法や平均場近似では，x^mとθの事後分布$p(x^m,\theta|y^m)$を$q(x^m,\theta|y^m)=q(x^m|y^m)\,q(\theta|y^m)$のような簡単な積の形でかける分布で近似し，その結果$q(\theta|y^m)$を得るという方法をとる．その近似においてpとqのKullback Leibler距離と呼ばれる統計的尺度$Eq[\ln(q/p)]$の最小化を基準とする．Variational Bayes法では最小化を，変分法で求めた$q(x^m|y^m)$と$q(\theta|y^m)$の反復計算で実現する．平均場近似ではさらに成分ごとの掛け算で表されるような分布を仮定し，反復法で求める．いずれも決定論的にかつ少ない計算量で求めることができる．

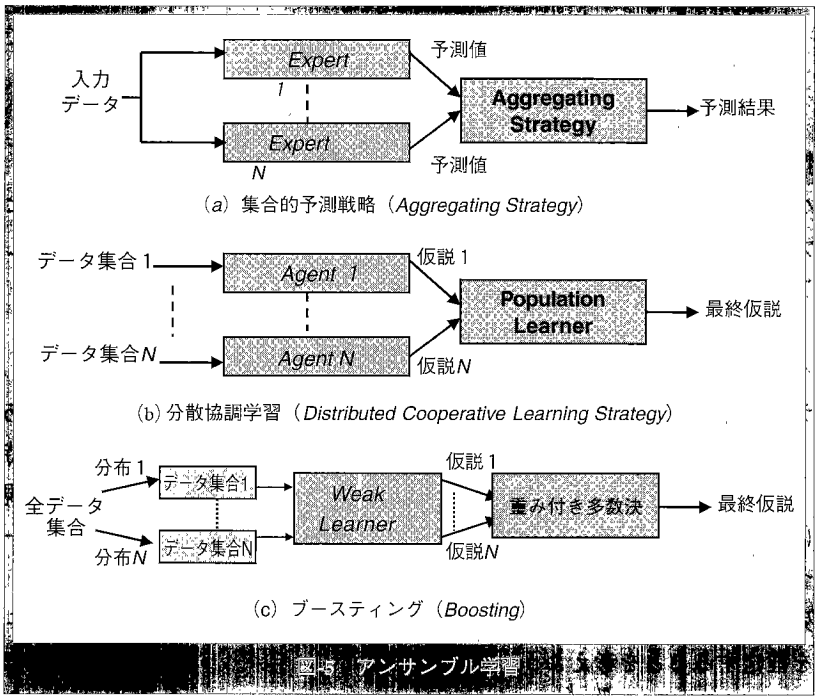


図-5 アンサンブル学習

■アンサンブル学習■

複数の学習アルゴリズムや学習のためのデータ収集機構を用意して，これらから寄せられた情報を統合して1つの学習結果を導き出す学習の方法はしばしば「アンサンブル学習」と呼ばれる．これは情報統合プロセスの問題として情報論的には大変興味深い問題を孕んでいる．ここでは以下の3つの形態を紹介する．

集合的予測戦略．逐次的予測問題において，いくつかの予測アルゴリズム（それぞれをExpertと呼ぶ）を用意して，これらの予測結果を結合することにより最終的な予測結果を出力する戦略が考えられる．これが**集合的予測戦略**（Aggregating Strategy）と呼ばれるものである（図-5 (a)）[6]．この戦略では，各Expertは同じ入力を与えられ，それらの予測結果を結合するとき，それらの重み付き平均に基づいて予測を行う．その際，これまでの予測損失が小さいExpertほど高い重みを割り付けるようにするが，この重みがパーセプトロンのアルゴリズムのように加減算で更新されるのではなく，乗算的に更新していくところが特徴である．このような戦略によって，最良のExpert（これは事前には分からない）が単独で予測したときの損失に最も近づけることができることが知られている．集合的予測戦略は統計学のベイズ混合予測的な考え方を一般の損失関数をも扱える形に一般化したものである．これは逆に統計学にも刺激を与えている．

分散協調学習．一括学習において，データを集める端末が分散している場合，それぞれの端末sublearner（これをAgentと呼ぶ）が学習結果としての仮説をセンタに送り，センタのPopulation learnerが最終的な学習結果を出力するような戦略が考えられる．これが**分散協調学習**（Distributed Cooperative Learning Strategy）と呼ばれるものである（図-5 (b)）[8]．

各Agentには異なる入力が与えられるが，Agentがそのままデータを Population learnerに渡してしまえばただの一括学習と同じになってしまう．そこで，分散協調学習の目的は各Agentが並列分散的に学習を行い，圧縮された学習結果をPopulation learnerに渡すことによって全体的な学習のスピードアップを図ることである．そこで，できるだけ情報を損失せずに最大限にスピードアップできるためにどのように学習を行えばよいか？　分散端末数はどれくらいに設定すればよいか？が問題となる．これ対しては，多端子情報理論的な考え方が解答を与えてくれるが，統計物理学のアルゴリズムや並列計算の理論とも結び付いて総合的な問題へと発展している．

ブースティング．一括学習において，与えられたデータ集合からリサンプリングを繰り返して複数回学習を行い，そこで得られた複数の仮説の重み付き多数決で決まる関数を最終仮説として出力する戦略を**ブースティング**(Boosting)と呼ぶ(図-5(c))[3]．ここで，仮説生成用に用いられる学習アルゴリズムをWeakLearnerと呼ぶ．最初は与えられたデータ集合上の確率分布を一様分布としておき，WeakLearnerがこの分布に従ってリサンプリングされたデータから学習を行い，そこで得られた仮説が与えられた全データに対してどれだけ誤って予測するかを計算して，誤ったデータに対しては確率値を増し，正しく予測できたデータについては確率値を減らしながらデータ集合上の確率分布を更新する．この仮説生成と確率分布の更新を複数回繰り返す．最終仮説はWeakLearnerよりも高い予測精度を持つことが示されている．リサンプリングにより仮説を生成するという点が統計学のBootstrap法にきわめて近い考え方を持つが，アンサンブル学習としての視点が逆に統計学にも刺激を与えている．

■その他の学習理論■

情報論的学習理論を舞台として展開されてきた学習の重要かつ新しい理論展開の一部を紹介しよう．

1）非正則な場合の学習理論：多層ニューラルネットや前出の混合モデル等のモデルは本来**非正則**である．すなわち，真のモデルよりも大きな自由度を持つ場合，真のパラメータ集合は1点ではなく，特異点を持った代数多様体の構造を持つ．このような場合，最尤推定量は存在せず，正則な場合の漸近評価は成り立たない．よってMDLやAICな統計的モデル選択基準は直ちに適用することはできないのである．近年，このような非正則なモデルを真っ向から取り上げた理論が発展している[14]．

2）能動学習と量子推定理論：与えられた入出力データからその間に成立する条件付確率分布を学習する際，入力分布を選択しながら学習を行うことを**能動学習**または適応型実験計画と呼ぶ．林と松本[9]は量子力学に従う系での**量子推定理論**の枠組みで能動学習を捉える新しいパラダイムを提案している．量子推定では未知状態の推定のみならずデータの抽出過程としての量子測定の仕方も問題にするが，これが入力分布の選択に相当するのである．

■応　用■

これまで紹介した学習技術は，自然言語処理，音声／画像処理，生体／遺伝子情報処理，パタン認識，金融工学，データマイニング，ロボティクスなどといった広範な分野に応用されている．特に自然言語処理では，語彙意味知識獲得やテキスト分類の問題において統計的モデル選択やアンサンブル学習の有効性が実証されている．画像処理では，セグメンテーション，幾何学的図形の認識問題に統計的モデル選択が適用されており，画像復元では統計力学的な最適化が威力を発揮している．以上の応用例は文献10）～15）にみることができる．

■おわりに■

情報論的学習理論の現状について大雑把に紹介した．興味深いトピックスで，ここに紹介しきれなかったものはたくさんある．たとえば，Graphical Model学習，独立成分解析や状態空間法といった時系列解析など．いずれの学習問題も「情報を最大限に抽出するためのプロセス」としての情報論的視点から捉えることができるであろう．しかし，本稿で紹介したトピックスはすべて情報論的学習理論という枠の中に素直に位置付けられるものでは決してない．個々の技術は，それぞれが接近しつつも微妙な緊張を保ちながら独自の展開の道を歩んでいる．繰り返しになるが，情報論的学習理論は，そこに共通する方法論について垣根を越えて議論をするための場を提供しているに過ぎない．しかし，このような学際的な視点は今後ますます重要になると思われる．

参考文献

1) Akaike, H.: A New Look at the Statistical Model Identification, IEEE Trans. Autom. Contr., AC-19, pp.716-723 (1974).
2) Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum Likelihood from Incomplete Data Via the EM Algorithm, J.Royal Statist. Soc. B, Vol.39, pp.1-38 (1977).
3) Freund, Y. and Schapire, R.: A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting, Jr.Computer and System Sciences, Vol.55, pp.119-139 (1997).
4) Rissanen, J.: Fisher Information and Stochastic Complexity, IEEE Trans. Inf. Theory, IT-42, 1, pp.40-47 (1996).
5) Shibata, R.: An Optimal Selection of Regression Variables, Biometrika, Vol.68, pp.45-54 (1981).
6) Vovk, V.G.: Aggregating Strategies, Proceedings of the Third Annual Workshop on Computational Learning Theory, Morgan Kaufmann, pp.371-386 (1990).
7) Yamanishi, K.: A Decision-theoretic Extension of Stochastic Complexity and Its Application to Learning, IEEE Trans. on Inf. Theory, IT-44, pp.1424-1439 (1998).
8) Yamanishi, K.: Distributed Cooperative Bayesian Learning Strategies, Information and Computation, Vol.150, pp.22-56 (1998).
9) 林, 松本: 適応型実験計画での二種類のBahadur型限界, 電子情報通信学会和文論文誌, Vol.J83A, No.6, pp.629-638 (June 2000).
10) 特集: 情報理論・統計的手法と学習技術, 計測と制御, Vol.38 (1999).
11) 情報論的学習理論小特集, 電子情報通信学会論文誌和文誌, Vol.J83A, No.6 (June 2000).
12) 1998年情報論的学習理論ワークショップ (IBIS'98) 予稿集 (1998).
13) 1999年情報論的学習理論ワークショップ (IBIS'99) 予稿集 (1999).
14) 2000年情報論的学習理論ワークショップ (IBIS'2000) 予稿集 (2000).
15) 情報論的学習理論特集, 人工知能学会論文誌, 2001年2月オンライン公開予定 (2001).

(平成12年11月30日受付)