解　説　能動学習

3. ニューラルネットワークの能動学習

Active Learning in Neural Networks by Kenji FUKUMIZU (Ricoh Co., Ltd.).

福　水　健　次[1]

1 (株)リコー研究開発本部

1. はじめに

システム同定や関数近似のように，真のシステム（学習課題）の入出力関係をパラメトリックな学習機械によって推定する問題はさまざまな場面で現れる．このとき学習データを作成するには，入力データ（質問）を用意しそれを真のシステムに入れて教師データ（答え）を観測するという作業が必要になる．データの入力や教師データの観測のためのコストが大きいこともあるため，より少ないデータ数によって高い精度を出す方法は重要である．その1つとして学習データを採取する入力点を最適化することが考えられるが，これは学習者が自ら質問を形成する学習法であるので，能動学習の一種と考えられる（図-1）．

このような学習データ採取法は，統計学の中では最適実験設計などと呼ばれ，線形回帰モデルを中心として古くから研究が行われている[1]．本稿では，数理統計的な側面からみた能動学習法を概観し，とくにニューラルネットワークへ応用する際の特有の問題点とその解決への試みを述べる．

2. 数理統計的な学習の枠組み

まず本稿で考える問題の枠組みを示す（図-2）．学習の目標となっている真のシステムの入出力関係は $\mathbf{R}^L$（L次元ユークリッド空間）から $\mathbf{R}^M$ への未知関数 $\boldsymbol{f}(\boldsymbol{x})$ で定められるとする．本稿では主として連続値の関数を想定している．この未知関数を学習するために，入力データ $\boldsymbol{x}^{(\nu)}$ ($\nu=1, 2, \cdots$) を用意してこれに対する真のシステムからの出力（教師データ）$\boldsymbol{y}^{(\nu)}$ を観測する．統計的に扱うために，観測には不可避なノイズが含まれると考えてこれを確率変数 Z で表し，学習データは

$$\boldsymbol{y}=\boldsymbol{f}(\boldsymbol{x})+Z \qquad (1)$$

という確率的な規則に従うとする．ここで Z は平均0のガウス性ノイズで，分散は σ^2 で各成分ごとの相関がないと仮定する．すなわち，

$$Z\sim\frac{1}{(2\pi\sigma^2)^{M/2}}\exp\left(-\frac{1}{2\sigma^2}\|z\|^2\right). \qquad (2)$$

本稿で論じる能動学習は，(1)式に従う独立な有限個の学習データ $D_N=\{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \cdots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})\}$ による学習を行う際に，入力データ $X_N=\{\boldsymbol{x}^{(\nu)}\}_{\nu=1}^N$ をいかに設計するかという問題として定式化される．

未知関数 $\boldsymbol{f}(\boldsymbol{x})$ の推定には，パラメータ $\boldsymbol{\theta}$ をもった関数族 $\{\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})\}$ を用意し，データに適合する $\hat{\boldsymbol{\theta}}$ を推定量として求め，$\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{\theta}})$ をもって関数 $\boldsymbol{f}(\boldsymbol{x})$ の推定値とする．たとえば多層パーセプトロンでは，$\boldsymbol{\theta}$ は結合荷重や閾値をまとめて表したベクトルであり，

$$f^i(\boldsymbol{x};\theta)=\sum_{j=1}^{H}w_{ij}s\left(\sum_{k=1}^{L}u_{jk}x_k+\zeta_j\right)+\eta_i \qquad (3)$$

($s(t)=1/(1+e^{-t})$ はシグモイド関数）と書ける．また線形回帰モデルの場合には，

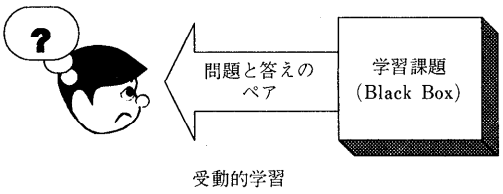
$$f(\boldsymbol{x};\boldsymbol{\theta})=\sum_{i=1}^{L}\theta_i x_i \qquad (4)$$

のように関数族を与える．推定量には以下の最小2乗誤差推定量を用いる．

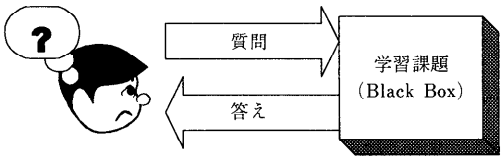
$$\hat{\boldsymbol{\theta}}=\arg\min_{\theta}\sum_{\nu=1}^{N}\|\boldsymbol{y}^{(\nu)}-\boldsymbol{f}(\boldsymbol{x}^{(\nu)};\boldsymbol{\theta})\|^2. \qquad (5)$$

ここで，arg min は最小値を達成するパラメータ $\boldsymbol{\theta}$ を表す．数理統計学的には $\hat{\boldsymbol{\theta}}$ は，学習データに対する分布のあてはまり度を表す尤度を最大化する最尤推定量に一致する．

本稿では最尤推定量に関する能動学習に焦点を絞る．そのほかに，パラメータ $\boldsymbol{\theta}$ も確率的な対象とみなして学習データに対する $\boldsymbol{\theta}$ の事後確率を求めるBayes推定の枠組みでの能動学習も研究されているが，その説明は文献2),3)などに譲

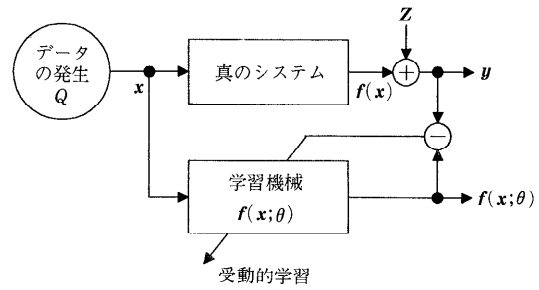


受動的学習

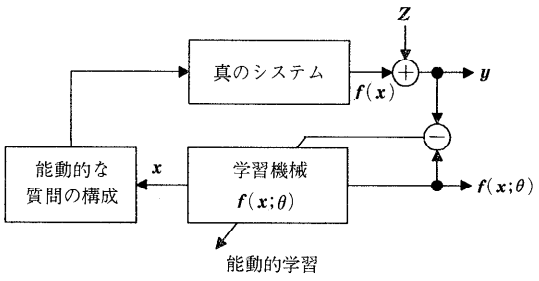


能動的学習

図-1 2種類の学習形態



受動的学習



能動的学習

図-2 数理統計的な学習の枠組み

る．文献2)の方法は導出にはBayes推定を用いているものの，正規近似の結果得られる基準は本稿のものと本質的に同じである．

能動学習では学習効率を議論するので，学習結果を評価する基準を導入すべきであろう．そのために，真のシステムはある一定の環境に設置されていて，分布 Q に従って確率的に発生する入力を受けとると仮定する．このとき，機械による予測値と真の出力値との平均2乗誤差

$$E(D_N)\equiv\int\|\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{\theta}})-\boldsymbol{f}(\boldsymbol{x})\|^2 Q(d\boldsymbol{x}) \quad (6)$$

を**予測誤差**と呼ぶ．学習データが(1)式の確率的規則に従う場合には，さらに期待値をとった

$$\mathcal{E}(X_N)=\mathrm{E}_{Y_N}\left[\int\|\boldsymbol{f}(\boldsymbol{x};\hat{\boldsymbol{\theta}})-\boldsymbol{f}(\boldsymbol{x})\|^2 Q(d\boldsymbol{x})\right] \quad (7)$$

を**期待予測誤差**と呼び学習結果の評価基準とする．

能動学習と対比して，学習データの入力を Q からのサンプルによってとるとき，この学習は**受動的学習**と呼ばれる．

以上の準備に基づくと，能動学習法の具体的な構成には次のような課題があることがわかる．

- 期待予測誤差をどのように計算するか？
- X_N をどのように最適化するか？

これらの課題に関する研究を以下に述べる．真の関数 $\boldsymbol{f}(\boldsymbol{x})$ は未知であるから，第1の問題には何らかの推定値を用いる必要がある．これには，一般に最尤推定量の漸近理論を用いる方法やBootstrapを用いる方法が考えられる[8]．しかしBootstrap法では，新たに学習データをとる点 $\boldsymbol{x}$ が予測誤差にどのように関係するかを知ることが困難であるため，漸近理論による近似がよく用いられる[4),5)]．

以下では真の関数 $\boldsymbol{f}(\boldsymbol{x})$ は設定したモデル $\{\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})\}$ に含まれるとし，$\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}_0)=\boldsymbol{f}(\boldsymbol{x})$ と仮定する．現実の問題でこれが完全に満足されることは少なく，理論導出のための理想的な仮定といえる．このことが能動学習を応用する際の1つの大きな問題となるが，それについては後述する．

漸近理論に従うと，学習データ数 N が大きいとき(7)式は以下のように近似できる．

$$\mathcal{E}(X_N)\approx\sigma^2\,\mathrm{Tr}[I(\boldsymbol{\theta}_0)J^{-1}(\boldsymbol{\theta}_0;X_N)]. \quad (8)$$

ここで行列 I,J はFisher情報行列と呼ばれ，

$$I_{ab}(\boldsymbol{x},\boldsymbol{\theta})=\frac{\partial\boldsymbol{f}^T(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_a}\frac{\partial\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_b},$$

$$I(\boldsymbol{\theta})=\int I(\boldsymbol{x};\boldsymbol{\theta})dQ(\boldsymbol{x}),$$

$$J(\boldsymbol{\theta};X_N)=\sum_{\nu=1}^{N}I(\boldsymbol{x}^{(\nu)};\boldsymbol{\theta}), \quad (9)$$

により与えられる．

(8)式は未知パラメータ $\boldsymbol{\theta}_0$ を含んでいるため，この値をそのまま用いることはできない．そこで

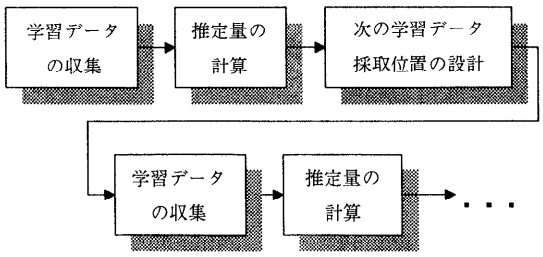


図-3 シーケンシャルな学習

逐次的に学習データを増やしていき，現在の推定量 $\hat{\boldsymbol{\theta}}$ で $\boldsymbol{\theta}_0$ を置き換えて(8)式を計算し，次の学習データ採取点を決定していくというシーケンシャルな方法をとる必要がある（**図-3**）．すなわち N 個までの学習データで学習した後，$N+1$ 個目の入力点を

$$\boldsymbol{x}_{N+1}=\arg\min_{\boldsymbol{x}} \mathrm{Tr}[I(\hat{\boldsymbol{\theta}}_N)J^{-1}(\hat{\boldsymbol{\theta}}_N ; X_N \cup \boldsymbol{x})] \tag{10}$$

により決める．またこれにより，X_N をまとめて最適化せずに逐次的に次のものだけを最適化すればよくなるので，最適化問題の次元削減にもなっている．なお，用意するモデルが線形モデルのときには上記の I, J はパラメータ $\boldsymbol{\theta}$ に依存せず，最適な X_N を直接求めることも可能である[6]．

よく知られているように，受動的学習においては，期待予測誤差をさらに X_N によって平均した値は，モデルによらず

$$E_0[\mathcal{E}(X_N)] \approx \frac{\sigma^2}{N} \times (\text{パラメータ数}) \tag{11}$$

と近似される．能動学習を行うことにより，この値よりも予測誤差を小さくすることが期待される．

3. 能動学習の方法

3.1 種々の能動学習法

これまでの考察により，能動学習の方法をどのように構成すればよいかほぼ明らかになった．しかしながら，ニューラルネットワークのような非線形モデルに適用するためには，なお考えなければならない問題がある．1つの問題はニューラルネットの学習が必ずしも最尤推定量を求めていない点にある．パラメータ $\boldsymbol{\theta}$ に関して非線形性なモデルでは最小2乗誤差推定量を直接解くことはできないので，バックプロパゲーションなど最急降下法的な手段で準最適解を求めることになるが，これらの方法は最小値ではなく局所解（極小値）を求めてしまう場合が多い．能動学習で学習データを最適化すると，より局所解に陥りやすい学習データを生成する可能性がある．実際，文献7)には中間素子が1個からなる簡単な多層パーセプトロンにおいて，データ採取点を最適化することによりかえって局所解に陥りやすくなる例が示されている．実は，パラメータ数が S のとき，Fisher 情報行列に基づく能動学習での最適学習データは，多くとも $S(S+1)/2$ 個の離散点上で採取すればよいことが知られているが[1]，このような学習データは局所解を導きやすい．したがってある程度ばらつきを残したデータ採取法を考えないと現実には応用できない．これに対して以下のような解決策が考えられる．

[**参照点を用いる方法**]

文献4)では(10)式をそのまま用いずに，学習データ採取ごとにばらつきをもつ参照用サンプル $\boldsymbol{x}_r$ をとり，

$$\boldsymbol{x}_{N+1}=\arg\min_{\boldsymbol{x}} \mathrm{Tr}[I(\boldsymbol{x}_r ; \hat{\boldsymbol{\theta}})J^{-1}(\hat{\boldsymbol{\theta}} ; X_N \cup \boldsymbol{x})] \tag{12}$$

により次の学習データ採取点を選択している．これは，$I(\hat{\boldsymbol{\theta}})$ の積分の計算負荷の問題を解決するだけでなく，$\boldsymbol{x}_r$ のバリエーションにより局所解の問題を回避していると考えることができるが，Q により積分した(8)式の予測誤差最小基準とは一致しないことに注意されたい．

[**確率的能動学習**]

文献5)では，確率的なデータ採取法によって局所解の問題に対処している．データを採取するための確率密度関数族 $\{r(\boldsymbol{x} ; \boldsymbol{v})\}$（$\boldsymbol{v}$ はパラメータ）を用意し，(8)式を X_N に関して平均した値

$$\mathrm{E}_{X_N}[\mathcal{E}(X_N)]=\frac{\sigma^2}{N}\mathrm{Tr}[I(\boldsymbol{\theta}_0)J^{-1}(\boldsymbol{\theta}_0 ; \boldsymbol{v})] \tag{13}$$

を最小にする分布 r を学習データの発生に用いる．ここで，

$$J(\boldsymbol{\theta} ; \boldsymbol{v})=\int I(\boldsymbol{x} ; \boldsymbol{\theta})r(\boldsymbol{x} ; \boldsymbol{v})d\boldsymbol{x} \tag{14}$$

である．具体的な方法を**図-4**に示す．ここで $N_0+N_1+\cdots+N_T=N$ とする．

この方法は確率的な最適性しかない上に確率分布族を勝手に設定するため，最適な分布が得られる保証がないが，データの採取位置がばらつくため局所解の回避に有効であると期待できる．

[多点探索を利用する方法]

逐次的に入力点を選択する(10)式を厳密に最大化するのではなく，$\boldsymbol{x}$ の候補をいくつか発生させて，その中で最大値を与えるものを $\boldsymbol{x}_{N+1}$ とすることが考えられる．この方法では候補点が多くなれば最適解に近づくので，N に応じて候補点を増加させれば，学習の初期には適当なばらつきをもったデータ設計が可能となり局所解が回避しやすいと考えられる．候補点を Q に従って発生させれば，データ採取点は Q 自身から出発して徐々に最適設計へと向うことになる．

3.2 他の基準に基づく能動学習

ここで予測誤差とは異なる基準による能動学習について簡単に触れておく．統計学の分野で最適実験設計として研究が盛んなのは $\det J(\boldsymbol{\theta}_0; X_N)$ を最大化する D-optimality と呼ばれる基準である．これは適当な条件のもとで Minimax 基準

$$\min_{X_N} \max_{x} \|\boldsymbol{f}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) - \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}_0)\| \qquad (15)$$

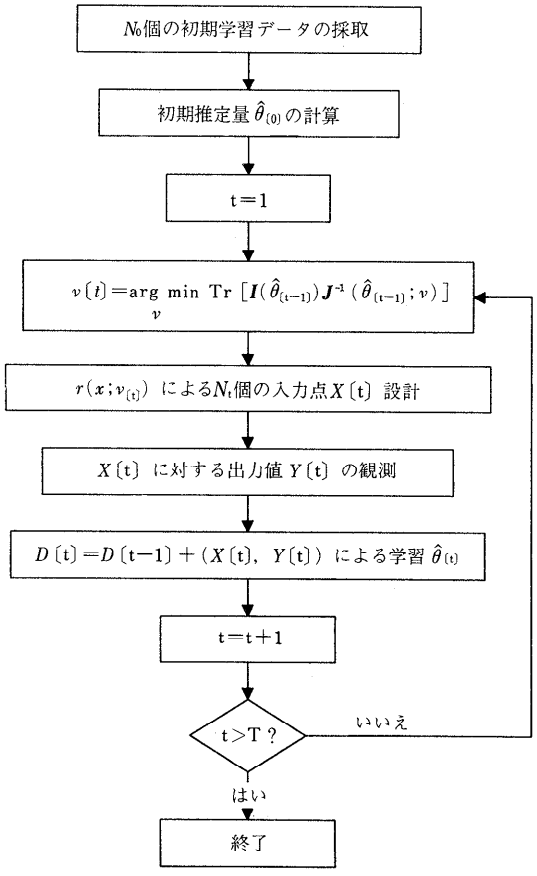


図-4 確率的能動学習

と一致することが知られている[1]．また文献3)にあるように各入力点 $\boldsymbol{x}$ における変動

$$V(\boldsymbol{x}) = \mathrm{E}[\|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}; \hat{\boldsymbol{\theta}})\|^2] \qquad (16)$$

の推定値を求め，$V(\boldsymbol{x})$ の最大値を与える $\boldsymbol{x}$ を次の学習データ採取点とする方法もある．この方法は D-optimality を漸近的に達成する設計法を与える[1]．

ここで触れた方法は，平均する分布 Q を仮定しておらず，平均2乗誤差に基づく期待予測誤差基準とは異なる．どれを能動学習の基準として用いるかは設計者の目的に依存している．

3.3 実験結果

ここでは文献5)に基づいて，能動学習を用いた簡単な実験結果を示す．確率的能動学習法，多点探索法，および(16)式の変動最大点を基準とする方法を3層パーセプトロン（MLP）に用いた簡単な実験結果を示す．この実験では，入出力は1次元で中間素子1個のモデルを用い，真の関数は $f(x)=s(x)$ とした．学習データは最終的に100個とし，それぞれ初期学習には10個のデータを受動的に作成した．Q としては平均0，分散1の正規分布を用い，(2)式の σ は0.01とした．

確率的能動学習では $r(\boldsymbol{x}; \boldsymbol{v})$ として4個の正規分布の混合分布を用い，10個ずつデータを発生させた．多点探索法では N 番目の採取点に対し候補点を N 個とった．30回の実験に対する予測誤差の平均を**図-5**に示す．能動学習が確かに受動的学習よりも誤差を小さくしているのがわかる．3つの能動学習のなかでは確率的能動学習が

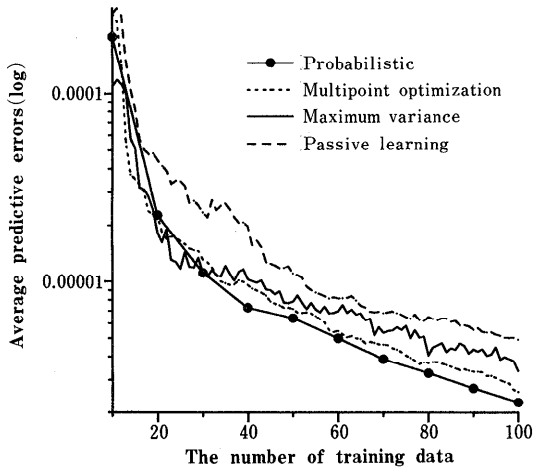


図-5 学習結果の比較

最も効果が高い．変動最大点法は予測誤差を基準としていないため学習データ数が多くなると予測誤差の減少はほかの2つに比べて小さいが，学習の初期には効率のよい学習を達成している．

3.4 モデル不適合の問題

能動学習法を導く際に真の関数が設定したモデルに含まれているという仮定をおいたが，現実の問題ではこの仮定が満されないことのほうが多い．モデルが真の関数を含んでいない場合でも予測誤差を推定する方法は知られているが[8]，推定精度がそれほどよくない上に，能動学習のように学習用のデータと評価用のデータが異なる分布に従う場合にはそのまま用いるのは困難である．

そこで，真の関数がモデルに含まれるという仮定がよい精度で実現できるように，モデルのサイズなどを適切に設定することが能動学習ではとくに重要となる．このため文献3)や5)ではモデル選択をしながら能動学習を行う方法を提案している．次節ではモデルを縮小させながら多層パーセプトロンの能動学習を行う方法を紹介する．

4. 多層パーセプトロンの能動学習

ここでは3層パーセプトロンに限定して，中間素子を削減しながら能動学習を行う方法を述べる．3層パーセプトロンでは，中間素子の個数を十分多くとるとコンパクト集合上の任意の連続関数がいくらでも精度よく近似できることが知られているため，十分大きいモデルを使えばモデル不適合の問題は回避できそうに思える．しかしながら，大きすぎる中間素子数はよく知られたパラメータ数の増加にともなう予測誤差の劣化を生じるだけでなく，Fisher情報行列の逆行列が存在しなくなり，これに基づく能動学習法が適用できないという問題を引き起こす[9]．この問題は多層パーセプトロンだけでなく一般に非線形モデルで生じ得る．

この可逆性の問題を解決するために，中間素子を削減しながら能動学習を行う方法が提案されている[5]．(3)式の記法に従うと，実は3層パーセプトロンにおいてFisher情報行列が非可逆になるのは，冗長な中間素子が存在する次の3つのケースに限られることが知られている[9]．

(1) $\exists j,\ \boldsymbol{u}_j := (u_{j1}, \cdots, u_{jL})^T = \boldsymbol{0}.$

(2) $\exists j,\ \boldsymbol{w}_j := (w_{1j}, \cdots, w_{Mj})^T = \boldsymbol{0}.$

(3) $\exists j_1 \neq j_2,\ (\boldsymbol{u}_{j1}, \zeta_{j1}) = \pm(\boldsymbol{u}_{j2}, \zeta_{j2}).$

上の3条件はそれぞれ，すべての入力素子からの結合が0の中間素子が存在する，すべての出力素子への結合が0の中間素子が存在する，2つの中間素子の役割を1個の中間素子で代用できる，という場合に対応している．そこで，この3ケースをチェックして，学習の最中に冗長と判断される中間素子が発生したらそれを削除するという手順を導入すれば，Fisher情報行列を可逆に保ち能動学習を適用可能にできる．

この方法を用いて，真の関数がモデルに含まれない場合の確率的能動学習の実験を行った結果を**図-6**に示す．$\boldsymbol{x}$ を4次元，$\boldsymbol{y}$ を1次元とし，真の関数は誤差関数erfを使って

$$f(\boldsymbol{x}) = \mathrm{erf}(x_1) \qquad (17)$$

により定めた．ここでerfは $\mathrm{erf}(x) =$

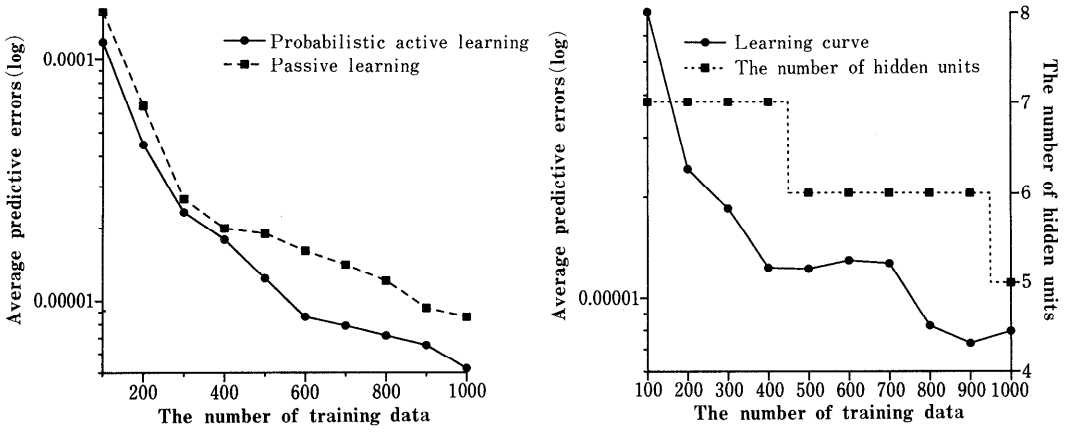


図-6 中間素子削減付の確率的能動学習

$\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ で定義される．誤差関数はシグモイド関数に形状は似ているが，多層パーセプトロンでは完全には実現できない．モデルは中間素子7個から出発した．図-6には平均の予測誤差と，ある学習過程における中間素子の削減の様子が示されている．中間素子が削減されることによりFisher情報行列を用いた能動学習が可能となっており，能動学習の効果も表れていることがわかる．

5. おわりに

ニューラルネットへの応用を念頭において，数理統計的な観点からみた能動学習法について述べた．本稿では2乗誤差で計算された予測誤差を小さくするような能動学習を主として論じた．とくにニューラルネットワークのような非線形モデルに能動学習法を用いる際には，局所解の問題やFisher情報行列の退化の問題を解決しなければならないことを説明した．非線形モデルではこの退化の問題により受動的学習以上にモデル選択が重要になることを述べ，多層パーセプトロンの場合にその解決方法の1つを示したが，モデルが多少不適合でも安定して効果を発揮する能動学習法は今後も重要な研究課題だと考える．

統計的な能動学習は，線形モデルを中心に長い研究の歴史をもっているが，現実の問題に応用されている例は比較的少ないように感じる．さらに，本稿では数理統計的に扱いやすい問題設定を行ったが，もっと動的な性質をもったシステムの学習に対してこそ「能動的」という言葉がふさわしいようにも思える．しかしながら，そのような場合の理論展開はあまり見あたらないようである．本稿を通して少しでも多くの読者が能動学習に興味をもち，具体的な問題への応用や異なった枠組みへの発展を試みていただければありがたく思う．

参考文献

1) Fedorov, V. V.: Theory of Optimal Experiments, Academic Press, New York (1972).
2) MacKay, D.: Information-based Objective Functions for Active Data Selection, Neural Computation, Vol. 4, No. 4, pp. 305-318 (1992).
3) Kindermann, J., Paass, G. and Weber, F.: Query Construction for Neural Networks using the Bootstrap, Proc. Intern. Conf. Artificial Neural Networks 95, pp. 135-140 (1995).
4) Cohn, D. A.: Neural Network Exploration Using Optimal Experiment Design, Advances in Neural Information Processing Systems 6 (Cowan, J. et al. (ed.)), Morgan Kaufmann, San Meteo, pp. 679-686 (1994).
5) Fukumizu, K.: Active Learning in Multilayer Perceptrons, Advances in Neural Information Processing Systems 8 (Touretzky, D. et al. (ed.)), MIT Press, Cambridge, pp. 295-301 (1996).
6) 福水：多項式近似における学習データの最適設計と予測誤差，信学論A, Vol. J 79-A, No. 5, pp. 1100-1108 (1996).
7) 福水：能動学習－最適な質問の効果と問題点－，日本神経回路学会第7回全国大会講演論文集, pp. 153-157 (1996).
8) Tibshirani, R.: A Comparison of Some Error Estimates for Neural Network Models, Tech. Report, Dept. of Statistics, University of Toronto (1995).
9) Fukumizu, K.: A Regularity Condition of the Information Matrix of a Multilayer Perceptron Network, Neural Networks, Vol. 9, No. 5, pp. 871-879 (1996).

(平成9年4月30日受付)

福水 健次

1966年生．1989年京都大学理学部卒業．同年(株)リコー入社．現在，研究開発本部情報通信研究所に所属．ニューラルネットワーク，統計的学習の研究に従事．京都大学博士(理学)．日本神経回路学会，電子情報通信学会，Institute of Mathematical Statistics各会員．