論　文

階層型ニューラルネットワークの混合モデルによるベイズ最適な予測について

橋川　弘紀[†]　　後藤　正幸[††]　　俵　　信彦[†††]

On Bayes Optimal Prediction Based on Mixture Model of Multilayer Neural Networks

Hiroki HASHIKAWA[†], Masayuki GOTOH[††], and Nobuhiko TAWARA[†††]

あらまし　確率モデルの学習問題において，学習データと同じ母集団のデータ（未学習データ）の出力を精度高く予測することが重要であり，モデル選択は一つの解決方法となっている．しかし，目的を未学習データの出力の予測と考えた場合，必ずしもモデルを限定する必要はなく，このとき要求されるのは，精度の高い予測を行うことである．このような予測を考慮した確率モデルを構築する方法として，ベイズ決定理論に基づいた学習理論が広く研究されている．本論文ではまず，候補である複数のモデルすべての混合モデルを用いて予測することがベイズ最適であることを示す．しかし，このベイズ最適を一般の確率モデルに対して厳密に計算しようとすると，パラメータ空間上の複雑な積分操作が必要になり，計算が不可能になってしまう．そこで，ラプラスの方法を用いて，この積分操作を排除し，漸近近似的に事後予測分布を計算することによる漸近ベイズ最適な予測法を提案し，ニューラルネットワークモデルへ適用してその有効性を検証する．

キーワード　ニューラルネットワーク，ベイズ決定理論，混合モデル，汎化能力

1. まえがき

階層型ニューラルネットワーク（NN）は，非線形関係を有するシステム同定に有効とされ，パターン認識，文字認識，音声認識，故障診断，プラント制御などに広く活用されている．しかし，階層型NNの問題点の一つとして，汎化能力が十分でない場合があり，その向上が望まれている[1]～[4]．

この汎化能力は中間層のユニット数，結合荷重の数と密接な関係があり，数が少なすぎると学習が収束せず，多すぎると学習データに対して近似を良くしても，未学習データに対しては良い近似をするとは限らない．

そこで，中間層のユニット数を決定する方法として，少ないユニット数からユニットを増やしながら学習を進める生成的学習，大きいユニット数からユニットを削除していく削除的学習がある[1],[2]．また，結合荷重の数を決定する方法として，NNの構築法であるBP学習の評価関数に，ネットワークの複雑さを制限するペナルティ項を付加し，不必要な結合荷重を削除する構造学習，Weight Decay法などがある[1],[2]．一方，モデル選択に用いる情報量規準をNNに適用し，複数のモデルから最適なモデルを選択する試みもなされている[1],[3],[4]．情報量規準として例えば，AIC，MDL，BICなどが挙げられる[6]～[10]．

以上に示したような中間層ユニット数を決定する削除的学習・生成的学習，複数のモデルの中から一つ選択する情報量規準による方法などは，汎化能力が高いと考えられるモデルを一つ選択する方法である．また，BP学習の評価関数にペナルティ項を付加する方法は，結合荷重を削除してモデル構造を決定するので，削除的学習ととらえることもできるが，汎化能力の高いモデルを得るには，この学習則を用いてもモデル選択を行う必要があることが明らかにされている[11]．従って，これらの方法も汎化能力の高いモデルを一つ選択

† (株) 菱化システム，東京都
Ryoka Systems Inc., 1-28-38 Shinkawa, Chuo-ku, Tokyo, 104 Japan
†† 早稲田大学理工学部，東京都
School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169 Japan
††† 武蔵工業大学，東京都
Department of Industrial Engineering and Management, Musashi Institute of Technology, 1-28-1 Tamazutsumi, Setagaya-ku, Tokyo, 158 Japan

する方法と考えることができる．しかし，一つのモデルを選択するということは，他のモデルの可能性を捨て去っていると考えることもできる．また，確率モデルの学習法を，その目的から考えた場合，必ずしもモデルを一つに限定する必要はない．特に，目的を未学習データの出力の予測と考えた場合，望まれることは予測精度が高い学習法である．このような予測を考慮した確率モデルを構築する方法として，ベイズ決定理論[12],[13]に基づいた研究が行われている[16],[19]．

ベイズ決定理論に基づく学習の立場に立った場合，候補である複数の確率モデルすべての混合モデルを使って予測することがベイズ最適となる[13]〜[19]．しかし，このベイズ最適を一般の確率モデルに対して厳密に計算しようとすると，パラメータ空間上の積分操作が必要になり，一般に計算量的な困難が生じてしまう．そこで，本論文では，ラプラスの方法[14],[22]を用いて，この積分操作を排除することを考え，漸近近似的に混合モデルの事後予測分布を計算することによるベイズ最適な予測法を提案する．更に，提案法をNNモデルに適用し，シミュレーション実験を通じて提案法の有効性を示す．

2. 準　　備

2.1 3層階層型NN

入力層の第 i ユニット入力，出力を x_i^I，y_i^I とし，中間層の第 j ユニットの入力，出力を x_j^H，y_j^H とし，出力層の入力，出力を x_k^O，y_k^O とする．入力層と中間層，中間層と出力層の結合荷重を w_{ik}^{IH}，w_{jk}^{HO} とする．また，中間層，出力層のしきい値を θ_j^H，θ_k^O とする．入力層，中間層，出力層のユニットの入出力関数を f_{in}，f_{hid}，f_{out} とする．本論文では f_{in}，f_{out} を線形関数，f_{hid} をシグモイド関数とする．

このとき，NNの出力層第kユニットの出力は

$$y_k^O = f_{out}\left(\sum_{j=1}^{H} w_{jk}^{HO} f_{hid}\left(\sum_{i=1}^{I} w_{ij}^{IH} f_{in}(x_i^I) + \theta_j^H\right) + \theta_k^O\right) \tag{1}$$

で与えられる．

本論文ではNNの学習法として，高速学習法である共役こう配法を用いたBP学習を用いる[20],[21]．

2.2 問題設定

本論文では多入力1出力を扱うこととし，入力ベクトル $x_i = (x_{i1}, \cdots, x_{iI}) \in R^I$ に対する出力を $y_i \in R^1$ とする．これら，n 組の入力ベクトルと出力を学習データとし，$(x^n, y^n) = (x_1, y_1) \cdots (x_n, y_n)$ で表す．学習データが与えられたもとで，未学習データの入力ベクトル x_{n+1} が与えられたときの出力 y_{n+1} を予測する問題ととらえることができる．モデル族 $\mathcal{M}$ に含まれるNNモデルを M_k とし，$\boldsymbol{\theta}_k$ は M_k によって定まる適当な次元のパラメータである．候補のNNモデルは T 個とする．

NNの適用範囲は1) 非線形回帰問題，2) パターン認識問題の二つに大別できる．まず，これらの問題について定義する．

[定義 1]（非線形回帰問題）y_i の生成機構をシステム固有の確定出力 $f(x_i)$ に確率変数の実現値として表される雑音成分 ϵ_i が重畳したものとし，その条件付き同時密度関数を $p^*(y_i|x_i)$ と表す．$f(\cdot)$ は真の入出力関数で定義域上で連続であるとする．

$$y_i = f(x_i) + \epsilon_i \tag{2}$$

$$\forall i, \epsilon_i \sim N\left(0, \sigma^2\right) \tag{3}$$

問題は (x^n, y^n) と x_{n+1} が与えられたもとで，y_{n+1} を予測することである．□

y_i の確率構造を，

$$y_i = f_{NF}(x_i, \boldsymbol{\theta}_k, M_k) + \epsilon_i \tag{4}$$

で近似することを考える．ここで，$f_{NF}(x_i, \boldsymbol{\theta}_k, M_k)$ はNNの出力で，$\boldsymbol{\theta}_k$ はモデル M_k の結合荷重およびしきい値である．式(4)において，M_k，$\boldsymbol{\theta}_k$ が定まれば，y^n の条件付き同時確率密度関数

$$p\left(y^n \middle| x^n, \boldsymbol{\theta}_k, M_k\right) = \prod_{i=1}^{n} p\left(y_i \middle| x_i, \boldsymbol{\theta}_k, M_k\right) \tag{5}$$

が定まる．従って，非線形回帰問題をNNに適用した際のモデルのゆう度は，

$$\begin{aligned} p(y^n|x^n, \boldsymbol{\theta}_k, M_k) &= \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}_k, M_k) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} p\left(y_i - f_{NF}(x_i, \boldsymbol{\theta}_k, M_k)\right)^2\right\} \end{aligned} \tag{6}$$

となる．

[定義 2] （パターン認識問題）出力 y_i としては，各クラスを代表する出力とし，0,1 の 2 値で与えられるものとする．このとき，学習後の NN モデルの出力層ユニットの出力 $f_{NP}(x_i,\boldsymbol{\theta}_k,M_k)$ は，入力ベクトル x_i のもとで出力 y_i が 1 である確率を近似していると考えられ，

$$p(y_i=1|x_i,\boldsymbol{\theta}_k,M_k)\approx f_{NP}(x_i,\boldsymbol{\theta}_k,M_k) \tag{7}$$

で与えられる [3],[15]．

出力 y_i が 0 である確率は，

$$p(y_i=0|x_i,\boldsymbol{\theta}_k,M_k)\approx 1-f_{NP}(x_i,\boldsymbol{\theta}_k,M_k) \tag{8}$$

となる．従って，出力が得られる確率は，

$$\begin{aligned}&p(y_i|x_i,\boldsymbol{\theta}_k,M_k)\\&\approx\left(f_{NP}(x_i,\boldsymbol{\theta}_k,M_k)\right)^{y_i}\\&\quad\cdot\left(1-f_{NP}(x_i,\boldsymbol{\theta}_k,M_k)\right)^{1-y_i}\end{aligned} \tag{9}$$

となる．問題は，(x^n,y^n) と x_{n+1} が与えられたもとで，y_{n+1} のクラスを予測することである． □

独立な n 個の学習データに対して，出力 $\{y_i|i=1,\cdots,n\}$ が得られるゆう度は，

$$\begin{aligned}&p(y^n|x^n,\boldsymbol{\theta}_k,M_k)\\&=\prod_{i=1}^{n}p(y_i|x_i,\boldsymbol{\theta}_k,M_k)\\&\approx\prod_{i=1}^{n}\left(f_{NP}(x_i,\boldsymbol{\theta}_k,M_k)\right)^{y_i}\\&\quad\cdot\left(1-f_{NP}(x_i,\boldsymbol{\theta}_k,M_k)\right)^{1-y_i}\end{aligned} \tag{10}$$

となる．

3. 漸近的にベイズ最適な予測法の導出

3.1 ベイズ決定理論に基づく定式化 [13]～[19]

次のような損失関数を定義する．

$$\begin{aligned}&d_1(y_{n+1},Ay(y_{n+1}|x_{n+1},x^n,y^n))\\&=\left(y_{n+1}-Ay(y_{n+1}|x_{n+1},x^n,y^n)\right)^2\end{aligned} \tag{11}$$

$$\begin{aligned}&d_2(y_{n+1},Ay(y_{n+1}|x_{n+1},x^n,y^n))\\&=\begin{cases}0 & y_{n+1}=Ay(y_{n+1}|x_{n+1},x^n,y^n)\\1 & y_{n+1}\neq Ay(y_{n+1}|x_{n+1},x^n,y^n)\end{cases}\end{aligned} \tag{12}$$

ここで，$Ay(y_{n+1}|x_{n+1},x^n,y^n)$ は決定関数，d_1 は非線形回帰問題に対する損失関数，d_2 はパターン認識問題に対する損失関数である．リスク関数 R は決定関数 $Ay(y_{n+1}|x_{n+1},x^n,y^n)$ を用いたときのモデルの期待損失であり，確率モデルの悪さを測る尺度である．

$$\begin{aligned}&R_1(\boldsymbol{\theta}_k,M_k)\\&=\int_{y_{n+1}}d_1(y_{n+1},Ay(y_{n+1}|x_{n+1},x^n,y^n))\\&\quad p(y_{n+1}|x_{n+1},\boldsymbol{\theta}_k,M_k)\,dy_{n+1}\end{aligned} \tag{13}$$

$$\begin{aligned}&R_2(\boldsymbol{\theta}_k,M_k)\\&=\sum_{y_{n+1}}d_2(y_{n+1},Ay(y_{n+1}|x_{n+1},x^n,y^n))\\&\quad p(y_{n+1}|x_{n+1},\boldsymbol{\theta}_k,M_k)\end{aligned} \tag{14}$$

すべての $\boldsymbol{\theta}_k$，M_k に対して，R_1，R_2 を最小化する決定関数は存在しない．そこで，ベイズ決定理論では事後確率で平均化した平均的なリスクを最小化する戦略をとる．この平均リスクをベイズリスクと言い，次式で定義される．

$$\begin{aligned}&BR_i=\sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k}R_i(\boldsymbol{\theta}_k,M_k)\,p(\boldsymbol{\theta}_k,M_k|x^n,y^n)\,d\boldsymbol{\theta}_k,\\&\quad i=1,2\end{aligned} \tag{15}$$

$p(y_{n+1}|x_{n+1},x^n,y^n)$ は，事後予測分布，または事後混合分布と呼ばれ，

$$\begin{aligned}&p(y_{n+1}|x_{n+1},x^n,y^n)\\&=\sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k}p(y_{n+1}|x_{n+1},\boldsymbol{\theta}_k,M_k)\\&\quad p(\boldsymbol{\theta}_k,M_k|x^n,y^n)\,d\boldsymbol{\theta}_k\end{aligned} \tag{16}$$

で与えられる．

ベイズリスク BR を最小にする決定関数 $Ay(y_{n+1}|x_{n+1},x^n,y^n)$ は，以下のように与えられる．

[補題 1] 損失関数 d_1 に対するベイズリスク BR_1 を最小にする決定関数 $Ay(y_{n+1}|x_{n+1},x^n,y^n)$ は，

$$\begin{aligned}&Ay(y_{n+1}|x_{n+1},x^n,y^n)\\&=\int y_{n+1}p(y_{n+1}|x_{n+1},x^n,y^n)\,dy_{n+1}\end{aligned} \tag{17}$$

で与えられる．

(証明) [13],[16]～[19] 参照． □

[補題 2] 損失関数 d_2 に対するベイズリスク BR_2 を最小にする決定関数 $Ay\left(y_{n+1}|x_{n+1},x^n,y^n\right)$ は,

$$Ay\left(y_{n+1}|x_{n+1},x^n,y^n\right)=\begin{cases}0, & p\left(y_{n+1}=0|x_{n+1},x^n,y^n\right)>1/2\\ 1, & p\left(y_{n+1}=1|x_{n+1},x^n,y^n\right)>1/2\end{cases} \tag{18}$$

で与えられる. $p\left(y_{n+1}=0|x_{n+1},x^n,y^n\right)=p(y_{n+1}=1|x_{n+1},x^n,y^n)=1/2$ のときは, ランダムなテストを行うことで, 決定する.

(証明) [13],[16]~[19] 参照. □

3.2 事後予測分布の漸近近似法

3.1 で示したように, 非線形回帰問題, パターン認識問題に対して, 事後予測分布 $p\left(y_{n+1}|x_{n+1},x^n,y^n\right)$ をもとに予測を行うことがベイズ最適である. しかし, 一般の確率モデルに対して, $p\left(y_{n+1}|x_{n+1},x^n,y^n\right)$ を厳密に計算すると, パラメータ空間上の積分操作が必要になり一般的に計算が困難となる. そこで, ラプラスの方法 [14],[22] を用いて, この積分を排除し, 漸近近似的に事後予測分布を得る方法を提案する. そのために以下の定理を示す.

[定理 1] 適当な正則条件 [22],[23] のもとで, 未学習データの入力ベクトル x_{n+1} が与えられたときの出力 y_{n+1} の事後予測分布 $p\left(y_{n+1}|x_{n+1},x^n,y^n\right)$ は, 漸近的に,

$$p\left(y_{n+1}|x_{n+1},x^n,y^n\right)=\sum_{k=1}^{T}H_k p\left(y_{n+1}|x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k\right) \tag{19}$$

$$H_k=\frac{p\left(y^n|x^n,\hat{\boldsymbol{\theta}}_k,M_k\right)p\left(M_k\right)\left(\frac{n}{2\pi}\right)^{-\frac{s_k}{2}}}{\sum_{k=1}^{T}p\left(y^n|x^n,\hat{\boldsymbol{\theta}}_k,M_k\right)p\left(M_k\right)\left(\frac{n}{2\pi}\right)^{-\frac{s_k}{2}}} \tag{20}$$

となる. ここで, $\hat{\boldsymbol{\theta}}_k$ は最ゆう推定量, s_k は確率モデル M_k のパラメータ数, Z は規準化定数である.

(概証明) 付録参照. □

この事後予測分布を用いれば, 式 (17), (18) によりベイズ最適な予測が可能となる. 事後予測分布は複数のモデルの重み付け平均で表されるので, 混合モデルとも呼ばれる. ここで H_k に $p\left(y^n|x^n,\hat{\boldsymbol{\theta}}_k,M_k\right)$ と $n^{-\frac{s_k}{2}}$ が含まれることが重要である. $p\left(y^n|x^n,\hat{\boldsymbol{\theta}}_k,M_k\right)$ はデータへのフィッティングの度合を, s_k はモデルの複雑さを表しているので, データへのフィッティングがよく, パラメータ数の小さいモデルほど, 大きい重みが与えられることになる. 一般にこれらはトレードオフの関係にあり, H_k が最大になるバランス点が存在する.

4. 階層型 NN モデルへの適用

定理 1 の適当な正則条件とは, パラメータの事後分布が漸近正規性をもつための十分条件とほぼ等価である [22]. これには Fisher 情報行列の存在の仮定が含まれているが, ニューラルネットワークモデルでは Fisher 情報行列が縮退することが知られている [5],[11],[24]. 従って, この漸近式は厳密にニューラルネットワークには適用できない. しかし, 通常の情報量基準も同様の問題を抱えているが, 近似的, 形式的に適用するといった操作が行われる [5]. 本論文でも同様に, 形式的に漸近式をニューラルネットワークに適用する. 以上の理由から, 混合の重み付け方法としては, よりニューラルネットワークに合った方法が存在する可能性があり, これは今後の課題である.

4.1 非線形回帰問題の場合

式 (19) に式 (6) を代入すると, 非線形回帰問題に NN を適用した際の事後予測分布の漸近式は,

$$p\left(y_{n+1}|x_{n+1},x^n,y^n\right)=\sum_{k=1}^{T}H_k p\left(y_{n+1}|x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k\right)$$

$$H_k=\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-f_{NF}\left(x_i,\hat{\boldsymbol{\theta}}_k,M_k\right)\right)^2\right\}\cdot p\left(M_k\right)\left(\frac{n}{2\pi}\right)^{-\frac{s_k}{2}}\Big/Z\right] \tag{21}$$

で与えられる. ここで, $\hat{\boldsymbol{\theta}}_k$ は学習後得られた結合荷重およびしきい値である. s_k は NN のパラメータ数である. また, $p(M_k)$ はモデルの事前分布であり, Z は規準化定数である.

また, 補題 1 の決定関数に式 (21) を代入することにより, 損失関数 d_1 に対するベイズリスク BR を漸近近似的に最小にする決定関数 $Ay\left(y_{n+1}|x_{n+1},x^n,y^n\right)$

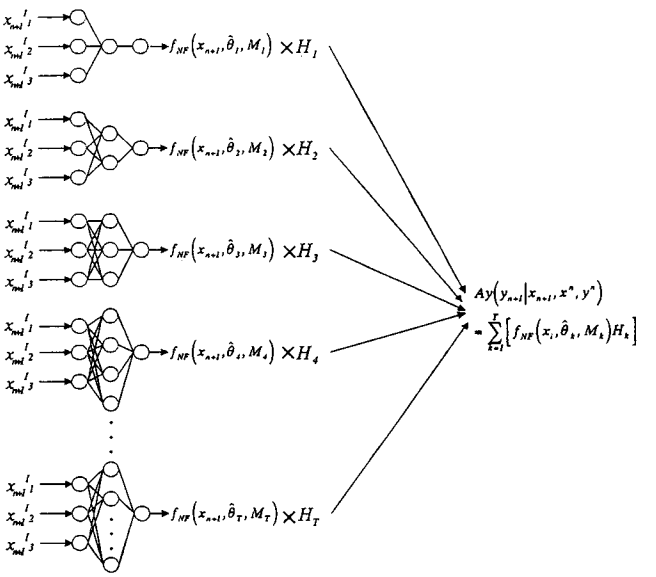

図1　非線形回帰問題の予測法
Fig.1　Prediction method for nonlinear regression.

は，

$$Ay\left(y_{n+1}|x_{n+1},x^n,y^n\right)=\sum_{k=1}^{T}H_k f_{NF}\left(x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k\right) \tag{22}$$

で与えられる．ここで，$\hat{\boldsymbol{\theta}}_k$ は学習後得られた結合荷重およびしきい値である．$f_{NF}\left(x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k\right)$ は NN モデル M_k に未学習データの入力ベクトル x_{n+1} を入力したときの出力である．

決定関数は，学習後の NN モデル M_k に未学習データの入力ベクトル x_{n+1} を入力したときの出力 $f_{NF}\left(x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k\right)$ の H_k による重み付き和で与えられる．H_k は漸近近似的にモデルの事後確率 $p\left(M_k|x^n,y^n\right)$ を表しており，ベイズ最適という意味付けが可能となる．但し，後でも考察されるように，Fisher 情報行列が退化するニューラルネットワークモデルに対しては，より適切な重み H_k の決定方法が存在する可能性がある．

非線形回帰問題に NN を適用した際の計算方法を図1に示しておく．

4.2　パターン認識問題の場合

式 (19) に式 (10) を代入すると，パターン認識問題に NN を適用した際の事後予測分布の漸近式は，

$$\begin{aligned}p\left(y_{n+1}|x_{n+1},x^n,y^n\right)&\approx\sum_{k=1}^{T}H_k\left(f_{NP}\left(x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k\right)\right)^{y_{n+1}}\\&\quad\cdot\left(1-f_{NP}\left(x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k\right)\right)^{1-y_{n+1}}\\H_k&=\left[\prod_{i=1}^{n}\left(f_{NP}\left(x_i,\hat{\boldsymbol{\theta}}_k,M_k\right)\right)^{y_i}\right.\\&\quad\left(1-f_{NP}\left(x_i,\hat{\boldsymbol{\theta}}_k,M_k\right)\right)^{1-y_i}\\&\quad\left.\cdot p\left(M_k\right)\left(\frac{n}{2\pi}\right)^{-\frac{s_k}{2}}\Big/Z\right]\end{aligned} \tag{23}$$

で与えられる．補題 2 より，ベイズ最適な予測は $p\left(y_{n+1}|x_{n+1},x^n,y^n\right)$ が最大となる y_{n+1} を予測値とする方法となる．

4.3　アンサンブル学習との比較

混合モデルを用いた研究としてアンサンブル（EL）学習 [25] がある．EL 学習では，主にタスクの異なるネットワークの混合モデルについて扱っている．これは学習データがある層別変数でいくつかの学習データセットに層別できる際に，これらが基本的構造を共有しながら，かつ，他と異なる構造も併せもつことを前提として，各データセットを別々のネットワークに学習させて最後に混合をとるモデルを構築している．EL 学習の本質的な目的は，層別されたの各データセットをそれぞれ別々のネットワークに学習させることによ

り，異なる構造を各ネットに別々に取り込み，共通にもつ構造は最後にすべてのネットワークの出力を平均化する部分で取り込むという形にモデルを限定することで，正しい構造を抽出しようとするものである．すなわち，このような学習データの背後に存在する構造に対する先験的知識をうまく使った結果が，EL 学習の形になったと言える．当然，各学習データを学習させるネットワークの複雑さなどは考慮されていない．これに対し，本論文の対象としている問題は，このような学習データの構造を考慮しているわけではなく，あくまで一組の学習データから各々のモデルが真である確率を計算しており，従来から統計的モデル選択で議論されてきた学習データ数と構造の複雑さのトレードオフを議論したものである．しかも，混合モデルによる予測方法が先にありきではなく，ベイズ最適な予測法を議論した結果が混合モデルとなったわけであり，ベイズ最適という理論的意味付けが可能である．すなわち，本論文の提案法は，従来法のように“混合モデルで予測することが有利な問題”を対象としているのではなく，一般的な問題に対してもベイズ決定理論に基づけばベイズ的な混合モデルが最適となることを示している．

5. シミュレーション

本章では非線形回帰問題を取り上げてシミュレーション実験を行い，提案方法の特性を考察する．

5.1 汎化能力の評価基準

非線形回帰問題においては，未来のデータの出力と NN の出力との差が小さければ予測精度が高いと考えられる．本論文では，学習データと同一母集団から発生している未学習データを評価用に用い，複数個の未学習データの出力と NN の予測出力との差をとり，これらの絶対値（絶対誤差）の平均と分散を評価関数とする．

5.2 学習データと評価用データ

対象モデルとして次式のような関数を用いる．

$$y = \sin(x_1) + \sin\left(\frac{x_2}{2}\right) + \sin\left(\frac{x_3}{3}\right) + \epsilon \qquad (24)$$

入力変数 (x_1, x_2, x_3) は $N(0, 1^2)$ に従う乱数を発生させ，それに誤差として $N(0, 0.1^2)$ に従う乱数を付加したモデルで実験を行う．学習データ数は 10,50,400 個の 3 種類を用意した．未学習の評価用データの入力変数は学習データと同様に 100 個発生させ，出力には誤差を付加しないものとした．

5.3 シミュレーション条件

用意した NN モデルは，入力層ユニット数が 3，出力層ユニット数は 1 で，中間層ユニット数が 1 個のモデルから 10 個のモデルの合計 10 個の NN モデルを用意した．結合荷重の初期値は -1.0〜1.0 の範囲で，各々のモデルに対し 10 通り乱数発生して学習させ，学習が最も収束したウェイトを採用した．また，学習の終了基準としては，2 乗誤差がほぼ収束するウェイトの更新回数 100 回と設定した．

5.4 シミュレーション結果と考察

図 2，図 3 および図 4 は，評価用データに対する絶対誤差の平均と分散を示している．混合モデルはすべてのモデルの重み付き平均なので，中間層素子数 1 から 10 の中で最も予測精度の高いモデルよりは，予測精度は必ず悪くなるが，その差が少なければ混合モデルの有効性が示される．すなわち，未学習の評価用データに対する予測精度が最も良いモデルを一つ，学習データから選択する方法はない．通常は AIC などにより，予測精度が高いと判断されるモデルを選ぶことになる．AIC との比較については次章でも示す．

図 2 は学習データが 10 個の場合の一例であり，学習データが少ないために予測精度がモデル間でかなりばらつき，規則性が見られない．混合モデルの予測精度は中間層素子数 4 のモデルとほぼ同程度となっている．AIC によって選択された中間層素子数はこの場合 9 となり，図 2 よりこのモデルが混合モデルより予測精度が劣ることがわかる．これは，データ数が少ないために AIC が導かれた漸近的状況が成り立っていないことによると思われる．混合モデルも同様の漸近論より導かれているが，全体を平均化する操作から，データ数が少ないことによる影響は少ないようである．

図 3 は学習データ数が 50 個の場合の一例で，予測絶対誤差の平均値は中間層素子数 4 で最小となり，それより大きいところで少し大きくなるという傾向が見られる．この場合も，混合モデルの予測精度は中間層素子 4 のモデルとそれほど変わりなく，誤差分散の方もほとんど同じになっており，予測精度は高いと言える．またこの場合，AIC は中間層素子数 8 を選択し，混合モデルより予測精度は劣っている．

図 4 は学習データ数が 400 個の場合の一例である．学習データが多くあるため，個々のモデルで中間層素子数に関係なくほぼ同じ結果となっている．本論文の予測方法はこれらの混合モデルで予測しているため，個々のモデルで予測したときの結果と変わらない．

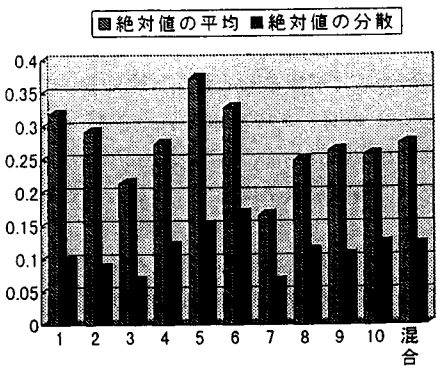

図2 学習データ数10個の結果
Fig. 2 Result on 10 learning data set.

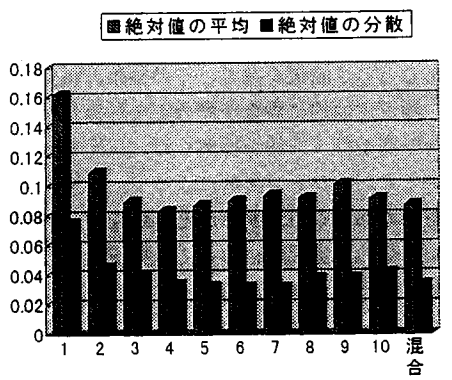

図3 学習データ数50個の結果
Fig. 3 Result on 50 learning data set.

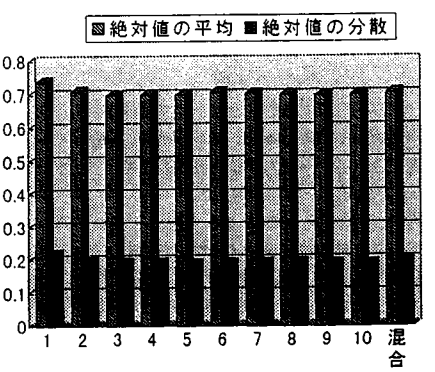

図4 学習データ数400個の結果
Fig. 4 Result on 400 learning data set.

以上により，本論文で提案した混合モデルでは，予測精度の高いモデルに大きい重みを割り当てており，最も予測精度の高いモデルに比較的近い予測が可能であることがうかがえる．従来のモデル選択の方法では，

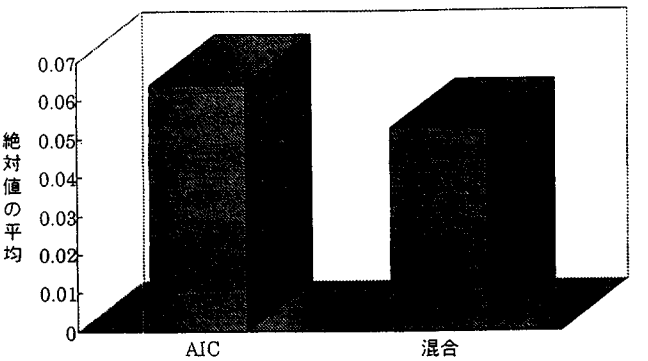

図5 絶対値の平均
Fig. 5 Absolute mean of prediction error.

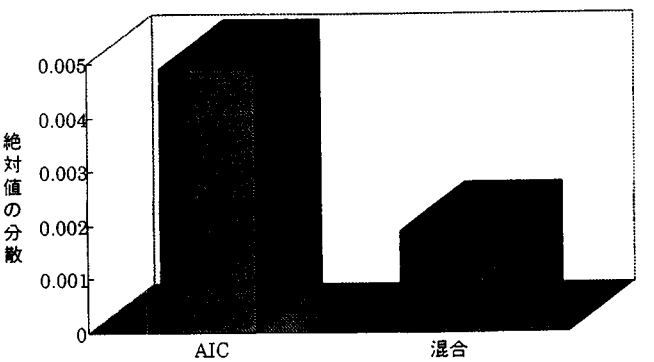

図6 絶対値の分散
Fig. 6 Variance of prediction error.

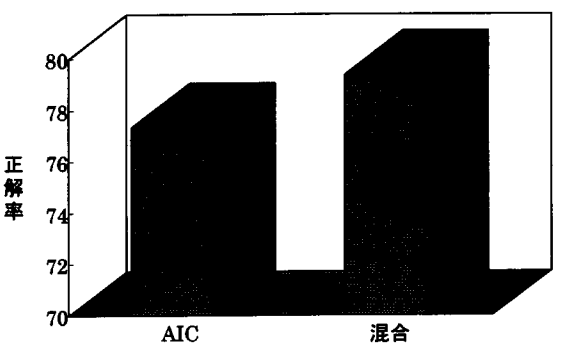

図7 正解率
Fig. 7 Accuracy rate of prediction.

最も予測精度の高いモデルを選択することはできないので，この結果は混合モデルの有効性を示していると言える．

5.5 実問題への適用

5.5.1 非線形回帰問題

対象データとして化学プロセスの制御問題を取り上

げた[14]．入力変数は8変数，出力には制御介入量をとり，学習データ数は350個とする．テスト用の未学習データは50個用意した．

（1） シミュレーション条件

用意したNNモデルは，入力層ユニット数が8，出力層ユニット数は1で，中間層ユニット数は1個のモデルから10個のモデルの合計10個を用意した．結合荷重の初期値は−1.0〜1.0の範囲で，各々のモデルに対し10通り乱数発生させて実験を行った．この中で最も学習が収束したモデルを用いた．総シミュレーション回数は100回である．また，学習の終了基準は学習回数が100回とした．

（2） シミュレーション結果

図5および図6の横軸は，情報量規準AICによって選ばれたモデル（中間層ユニットが5個のモデル）と，本論文で提案した混合モデルを用いる方法である．図5は，絶対値の平均を表していて，AICによる方法より，小さくなっている．AICで選択されたモデルを用いて予測した場合よりも平均値を約20%減少させることができた．そのことにより，50個の未学習データに対して平均的によく推定していると考えられる．図6は，絶対値の分散を表していて，小さくなっているので，50個の未学習データを均等に予測していると言える．AICのモデルを用いて予測した場合よりも分散を約60%減少させることができ，かなり均一に精度の高い予測を行っていると考えられる．

5.5.2 パターン認識問題

5.5.1の出力である制御介入量を以下のように分類した．正の介入を行ったら1，負であったら0とした．学習データ数は50個，テスト用の未学習データは150個用意した．

（1） シミュレーション条件

5.5.1のシミュレーション条件と同様である．

（2） シミュレーション結果

図7は正解率を表した図である．AICによって選ばれたモデルは中間層ユニット数が3個のモデルであり，未学習データ150個に対し，34個の誤った出力をした．一方，混合モデルを用いる方法は，31個の誤った出力をした．少しではあるが，本論文で提案した方法が良い結果となっている．

5.6 考　察

混合モデルを用いる予測法を非線形回帰問題のデータに対しシミュレーションを行った結果，この際に用いる事後予測分布の漸近近似式を用いた予測法は，有限のサンプル数に対しても有効であることが明らかとなった．また，この漸近的予測分布は，Fisher情報行列の存在を仮定して導出されるが，NNにおいては仮定できないことが明らかにされている[24]．従って，本論文ではこれを無視して，形式的に適用したことになるが，従来の情報量規準に基づく方法も同様の問題を抱えており[5],[11],[26]，今後の課題となっている．本論文では，学習が最も収束したモデルのパラメータを用いて事後予測分布の漸近近似式を計算しているが，シミュレーション実験によりその有効性を確認できたと思われる．

6. むすび

ベイズ決定理論に基づき，候補である複数のNNモデルすべての混合モデルを用いて予測することがベイズ最適であるという結果を導き，その結果の積分操作を削除し，漸近近似的に事後予測分布を計算して，ベイズ最適な予測を行う方法を提案した．NNモデルに適用し，シミュレーション実験によって，本論文の提案方法が有効であることが明らかとなった．

中間層素子数の決定問題においても指摘されているように，NNモデルに対しては一般の漸近論は厳密には適用できず，より理論的に適合する混合重みの決定法が議論できる可能性があり，これは今後の課題とする．

謝辞 本論文を精読し，有益なコメントおよびELに関するアドバイスを頂きました査読者の方に深く感謝致します．

文　献

[1] 喜多　一，“ニューラルネットワークの汎化能力，”システム/制御/情報，vol.36, no.10, pp.625–633, 1992.
[2] 石川真澄，“ネットワーク学習の最近の話題，”計測と制御，vol.30, no.4, pp.285–290, 1991.
[3] 栗田多喜夫，“情報量基準による3層ニューラルネットの隠れ層ユニット数の決定法，”信学論（D-II），vol.J73-D-II, no.11, pp.1872–1878, Nov. 1990.
[4] 栗田多喜夫，木村陽一，“階層型ニューラルネットワークとその周辺，”応用統計学，vol.22, pp.99–115, 1993.
[5] 渡辺澄夫，“情報量基準の変形によるニューラルネット最適化の一手法，”信学技報，NC93-52, 1993.
[6] 竹内　啓編，“統計学事典，”東洋経済新報社，1989.
[7] 坂本慶行，石黒真木夫，北川源四郎，“情報量統計学，”共立出版，1983.
[8] C. Schwarz, “Estimating the dimension of a model,” Annals of Statistics, vol.6, pp.461–464, 1978.
[9] D.S. Poskitt, “A Bayes procedure for the identification of univariate time series models,” The Annals of Statistics, vol.14, no.2, pp.502–516, 1986.

[10] D.S. Poskitt, "Precision, complexity and Bayesian model determination," J.R. Statist. Soc.B, vol.49, no.2, pp.199–208, 1987.
[11] 渡辺澄夫，福水健次，"縮退したFisher情報行列を持つ系の学習について，" 信学技報，NC94-56, 1994.
[12] J.O. Berger, "Statistical Decision Theory and Bayesian Analysis," Springer-Verlag, New York Heidelberg Berlin, 1985.
[13] T.S. Ferguson, "Mathematical Statistics—A Decision Theoretic Approach," Academic Press, New York, 1967.
[14] 韓　太舜，小林欣吾，岩波講座 応用数学 13 [対象 11]，"情報と符号化の数理，" 岩波書店，1994.
[15] 船橋賢一，横山俊和，"階層型ニューラルネットワークとベイズ識別理論の関係，" 信学技報，NC93-7, 1993.
[16] T. Matsusima, H. Inazumi, and S. Hirasawa, "A class of distortionless codes designed by bayes decision theory," IEEE Trans. Information Theory, vol.37, pp.1288–1293, 1991.
[17] 松嶋敏泰，稲積宏誠，平澤茂一，"不確実性をもつ論理式の帰納推論に関する一考察，" 情処学論，vol.33, no.12, pp.1461–1475, 1992.
[18] 後藤正幸，松嶋敏泰，平澤茂一，"ベイズ決定理論に基づくデータ解析に関する一考察，" 第18回情報理論とその応用シンポジウム予稿集，pp.533–536, 1995.
[19] 後藤正幸，松嶋敏泰，平澤茂一，"MDL基準に基づく符号とベイズ符号の符号長に関する解析，" 信学技報，IT96-3, 1996.
[20] 後藤正幸，開沼泰隆，俵　信彦，"変傾共役勾配法によるBP学習の安定化と高速化，" 日本経営工学会誌，vol.46, no.2, pp.152–158, 1993.
[21] 斉藤和己，中野良平，"3層ニューラルネットワークにおける2階導関数を用いた学習アルゴリズムの高速化，" 信学技報，NC94-7, 1994.
[22] B.S. Clarke and A.R. Barron, "Information-theoretic asymptotics of Bayes Methods," IEEE Trans. Information Theory, vol.36, no.3, pp.453–471, 1990.
[23] B.S. Clarke and A.R. Barron, "Jeffreys' Prior is Asymptotically Least Favorable under Entropy Risk," JSPI 41, pp.37–60, 1994.
[24] 萩原克幸，戸田尚宏，臼井支朗，"階層型ニューラルネットワークにおける結合重みの非一意性とAIC，" 信学論(D-II)，vol.76-D-II, no.9, pp.2058–2065, Sept. 1993.
[25] V. Tresp and M. Taniguthi, "Combining estimators using non-constant weight functions," in Advances in Neural Information Processing Systems (NIPS), G. Tesauro, et al. ed., vol.7, pp.419–426, 1995.
[26] 小野田崇，"階層型ニューラルネットワークの情報量基準，" 人工知能学会誌，vol.11, no.4, pp.574–584, 1995.

付　　録

定理1の概証明

厳密な証明は[22]の手法とほぼ同様にできる．ここでは，略解を示す．

ベイズの定理より，

$$p(\boldsymbol{\theta}_k, M_k|x^n, y^n) \propto p(y^n|x^n, \boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k, M_k) \tag{A·1}$$

であるから，

$$\sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k} p(y_{n+1}|x_{n+1}, \boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k, M_k|x^n, y^n)d\boldsymbol{\theta}_k$$
$$\propto \sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k} p(y_{n+1}|x_{n+1}, \boldsymbol{\theta}_k, M_k)$$
$$p(y^n|x^n, \boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k, M_k)d\boldsymbol{\theta}_k \tag{A·2}$$

となる．

式(A·2)を次のように変形する．

$$\sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k} p(y_{n+1}|x_{n+1}, \boldsymbol{\theta}_k, M_k)$$
$$p(y^n|x^n, \boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k, M_k)d\boldsymbol{\theta}_k$$
$$= \sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k} p(y_{n+1}|x_{n+1}, \boldsymbol{\theta}_k, M_k)$$
$$\exp\{\log p(y^n|x^n, \boldsymbol{\theta}_k, M_k)\}\, p(\boldsymbol{\theta}_k, M_k)d\boldsymbol{\theta}_k \tag{A·3}$$

式(A·3)の $\log p(y^n|x^n, \boldsymbol{\theta}_k, M_k)$ を最ゆう推定量 $\hat{\boldsymbol{\theta}}_k$ の周りでテイラー展開する．3次以下の項は確率収束の意味で $o(1)$ であることが示されるので，2次までを議論すると，

$$\sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k} p(y_{n+1}|x_{n+1}, \boldsymbol{\theta}_k, M_k)$$
$$\exp\{\log p(y^n|x^n, \boldsymbol{\theta}_k, M_k)\}\, p(\boldsymbol{\theta}_k, M_k)d\boldsymbol{\theta}_k$$
$$= \sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k} p(y_{n+1}|x_{n+1}, \boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k, M_k)$$
$$\exp\left\{\log p(y^n|x^n, \hat{\boldsymbol{\theta}}_k, M_k)\right.$$
$$\left.-\frac{n}{2}(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)\hat{I}(\hat{\boldsymbol{\theta}}_k)(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)\right\}d\boldsymbol{\theta}_k \tag{A·4}$$

ここで，

$\hat{I}(\hat{\boldsymbol{\theta}}_k)$

$$= \frac{1}{n}\left[\frac{\partial^2 \log p(y^n|x^n,\boldsymbol{\theta}_k,M_k)}{\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_k^T}\right]_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k} \quad \text{(A·5)}$$

である．

適当な正則条件によって，式(A·5)はFisher情報行列 $I(\hat{\boldsymbol{\theta}}_k)$ に確率収束する．また，$n\to\infty$ とすると，右辺の積分で実質的に効いてくるのは，$O(1/\sqrt{n})$ だけであるから，$\hat{I}(\hat{\boldsymbol{\theta}}_k)$，$p(y_{n+1}|x^n,\boldsymbol{\theta}_k,M_k)$，$p(\boldsymbol{\theta}_k,M_k)$ を $I(\hat{\boldsymbol{\theta}}_k)$，$p(y_{n+1}|x^n,\hat{\boldsymbol{\theta}}_k,M_k)$，$p(\hat{\boldsymbol{\theta}}_k,M_k)$ と置き換えると，

$$\begin{aligned}&\int_{\boldsymbol{\theta}_k} p(y_{n+1}|x_{n+1},\boldsymbol{\theta}_k,M_k)p(y^n|x^n,\boldsymbol{\theta}_k,M_k)\\&\quad\cdot p(\boldsymbol{\theta}_k,M_k)d\boldsymbol{\theta}_k\\&= p(y_{n+1}|x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k)p(y^n|x^n,\hat{\boldsymbol{\theta}}_k,M_k)\\&\quad\cdot p(\hat{\boldsymbol{\theta}}_k,M_k)\sqrt{\det I(\hat{\boldsymbol{\theta}}_k)}^{-1}\left(\frac{n}{2\pi}\right)^{-\frac{s_k}{2}} \quad \text{(A·6)}\end{aligned}$$

となる．s_k はモデル M_k のパラメータ数である．ここで，パラメータの事前分布としてJaffreys priorを仮定すると，

$$p(\hat{\boldsymbol{\theta}}_k|M_k)\propto\sqrt{\det I(\hat{\boldsymbol{\theta}}_k)} \quad \text{(A·7)}$$

であり，

$$\begin{aligned}&\sum_{k=1}^{T}\int_{\boldsymbol{\theta}_k} p(y_{n+1}|x_{n+1},\boldsymbol{\theta}_k,M_k)\\&\qquad p(y^n|x^n,\boldsymbol{\theta}_k,M_k)p(\boldsymbol{\theta}_k,M_k)d\boldsymbol{\theta}_k\\&\propto\sum_{k=1}^{T} p(y_{n+1}|x_{n+1},\hat{\boldsymbol{\theta}}_k,M_k)\\&\qquad p(y^n|x^n,\hat{\boldsymbol{\theta}}_k,M_k)p(M_k)\left(\frac{n}{2\pi}\right)^{-\frac{s_k}{2}} \quad \text{(A·8)}\end{aligned}$$

が得られる．□

(平成8年6月17日受付，11月19日再受付)

橘川　弘紀

平6武蔵工大・経営卒．平8同大大学院修士課程了．現在，(株) 菱化システム．在学中，ニューラルネットワークの統計的観点からの研究に従事．日本経営工学会会員．

後藤　正幸　(学生員)

平4武蔵工大・経営卒．平6同大大学院修士課程了．現在，早大・理工学専攻・博士後期課程在学中．理工学部・経営システム工学科・助手兼任．情報理論とその応用，統計的学習理論，統計的モデル選択，ベイズ統計，知的制御系などに興味をもつ．情報理論とその応用学会，日本経営工学会，人工知能学会，IEEE各会員．

俵　信彦　(正員)

昭34早大・理工・工業経営卒．昭36同大大学院修士課程了．現在，武蔵工大・経営工学科教授．工博．全社的品質管理，統計的品質管理，生産在庫管理，ニューラルネットワークおよび遺伝的アルゴリズムの理論とその応用などの研究に従事．平7より，日本経営工学会副会長．日本経営工学会，日本品質管理学会等各会員．