論　文

混合分布モデルと対数線形モデルに基づくフィードフォワード型ニューラルネット

正員　辻　敏夫[†]　非会員　市延　弘行[†]　非会員　金子　真[†]

A Proposal of the Feedforward Neural Network Based on the Gaussian Mixture Model and the Log-Linear Model

Toshio TSUJI[†], *Member*, Hiroyuki ICHINOBE[†] *and* Makoto KANEKO[†], *Nonmembers*

あらまし　本論文では，フィードフォワード型ニューラルネットの一般化能力を改善することを目的として，パターン識別問題によく用いられる混合正規分布モデルと対数線形モデルに基づいた新しいフィードフォワード型ニューラルネットを提案する．まず混合正規分布の各コンポーネントの事前確率と生起確率を一つにまとめ，これに対数線形モデルを適用することで事後確率を計算できることを示す．次にこの前向き計算をニューラルネットに展開し，ゆう度最大の評価のもとで学習則を導出する．本ネットワークは，(1)通常の誤差逆伝搬型ニューラルネットに入力データの分布を近似する統計構造を組み込んだ形になっているので，学習用データとして与えられていない入力データや少ない数の学習用データに対しても高い識別能力を実現できる，(2)混合正規分布モデルのパラメータを無制約化することで，より忠実に入力データの特性を表現できる，(3)統計手法を組み込むことで，ユニットの入出力関数，層の数，ユニットの数などのネットワーク構造を自然に決定することができる，(4)ニューラルネットからの出力値を確率として取り扱うことができるなどの特徴を有している．

キーワード　階層型ニューラルネット，混合正規分布，対数線形モデル，パターン識別

1. まえがき

誤差逆伝搬型ニューラルネット[(1)]は，高度に並列な処理が可能で，かつ任意の非線形写像を獲得できる強力な学習能力を有するため，現在までパターン識別や学習制御など多くの分野で用いられてきた．本論文ではこのニューラルネットの一般化能力を向上させることを目的とし，パターン識別問題によく用いられる混合分布モデルと対数線形モデルに基づいた新しいフィードフォワード型ニューラルネットを提案する．

一般に，パターン識別で用いられる入力データは，ある確率分布に従う確率変数である．この場合，入力データの事後確率を精度良く求めることができればパターン識別が実現できるので，パターン識別問題は通常，確率分布同定問題に帰着される．データが従う確率分布を推定する手法としては，あらかじめ特定の分布を仮定してそのパラメータを推定するパラメトリック法，たくさんのサンプルデータから度数分布を求め確率密度関数を推定するノンパラメトリック法がある．これらの手法をニューラルネットに応用した例として，例えばBridle[(2)]は，各事象に対して正規分布を仮定し，分布のパラメータをニューラルネットの重み係数として学習的に推定する方法を提案した．この方法は，少ないデータからパラメータを推定できるという特徴をもつが，仮定が成り立たないときには識別能力が大きく低下するというパラメトリック法の問題点をそのまま継承している．一方，中川・小野[(3)]はサンプルデータに対して，ベクトル量子化とRadial basis function networkを用いて確率密度関数と事後確率の推定を行った．このノンパラメトリックな方法も，データの分布を忠実に近似できる反面，サンプルデータを多く必要とするという問題点がある．

これらの方法に対して，分布のパラメータを有し，しかもデータに応じて分布の形状を変えられる性質をもつセミパラメトリック法[(4)]がある．母集団の分布を

† 広島大学工学部第二類，東広島市
Faculty of Engineering, Hiroshima University, Higashi-Hiroshima-shi, 724 Japan

有限個のコンポーネントと呼ばれる確率分布の線形和で近似する混合分布モデル(Mixture model)[5]はこのセミパラメトリック法の一つで，混合正規分布モデル(Gaussian mixture model)は，このコンポーネントに正規分布を用いたものである．この混合正規分布モデルとニューラルネットを組み合わせた方法として，Traven[4]，Perlovsky and McManus[6]，辻・森・伊藤[7]，Lee and Shimoji[8]らは，混合正規分布のパラメータを学習的に推定し，母集団の確率分布を精度良く近似できることを示した．しかしながらこれらの方法のほとんどは，各パラメータの推定量を計算するための繰返し法や確率分布を求めるための前向き計算をニューラルネットになぞらえて展開しているだけにすぎず，また混合正規分布を構成するコンポーネント数が増せば，分布を記述するパラメータの数も増大してしまう．

一方，Jordan and Jacobs[9]は一般化線形モデル[10]をニューラルネットに組み込んだHierarchical Mixture-of-Expert (HME)を提案した．一般化線形モデルとは，従属変数が指数分布族に従い，かつ，その期待値が入力ベクトルの線形結合に単調で微分可能な関数を作用した形で与えられる統計モデルのことで[10]，HMEではこの一般化線形モデルを用いてExpert networkとGating networkという二つの部分ネットワークを構成している．Expert networkは入力空間のある特定の領域におけるデータの特徴を抽出し，Gating networkはその特徴をどの程度，上位のユニットに伝えるかを調節することによって全体としての情報の統合化を行う．Jordan and Jacobsはパターン識別問題において，Gating networkとExpert networkの出力値がそれぞれ混合分布モデルの各コンポーネントの事前確率と生起確率に対応することを示した．このHMEでは，事前確率(Gating networkの出力値)が入力データに依存して変化するため，従来の混合正規分布モデルに比較してネットワークの表現能力は高い．しかしながら，その反面，Gating networkにもパラメータが必要となるため，混合正規分布モデルのもつパラメータ数よりはるかに多くのパラメータが必要となってしまう．彼らはこの問題に伴う学習の困難さを解決するため，Expectation-Maximizationアルゴリズム[11]に基づく複雑な学習則を提案している．

これらの研究に対して本論文では，混合正規分布の各コンポーネントの事前確率と生起確率を一つにまとめ，これに対数線形モデルを適用した新しいフィードフォワード型ニューラルネットを提案する．本ネットワークでは確率分布モデルをニューラルネットに組み込んでいるので，学習用データとして与えられていない入力データに対しても学習的に獲得した確率分布に基づいて事後確率を計算することができる．また，ユニットの入出力関数，層の数，ユニットの数など通常の誤差逆伝搬型ニューラルネットでは設定が困難なネットワーク構造も自然に決定することができる．以下，2.で対数線形モデルを混合正規分布モデルに組み込む方法，3.で本論文で提案するフィードフォワード型ニューラルネットの構造について説明し，4.で学習実験により本ネットワークの有効性を検討する．

2. 対数線形モデルを組み込んだ混合正規分布モデル

2.1 混合正規分布とパターン識別問題

混合正規分布モデルでは，それぞれ正規分布に従う有限個のコンポーネントを用い，その線形和で母集団の確率分布を表現する[5]．このとき，各コンポーネントのパラメータが推定できれば，母集団の分布を近似的に推定できたことになる．今，特徴ベクトル $x \in \mathcal{R}^d$ の生起確率密度関数を事象数 K (各事象 k のコンポーネント数は M_k とする；$k=1,\cdots,K$) の混合正規分布で表現する．各コンポーネント $\{k,m\}$ のもつパラメータは，混合度(mixing proportion[5]) $\alpha_{k,m}$，平均ベクトル $\mu^{(k,m)} \in \mathcal{R}^d$，共分散行列 $\Sigma^{(k,m)} \in \mathcal{R}^{d\times d}$ である．このとき，x の確率密度関数 $f(x)$ は，

$$f(x)=\sum_{k=1}^{K}\sum_{m=1}^{M_k}\alpha_{k,m}g(x;\mu^{(k,m)},\Sigma^{(k,m)}) \tag{1}$$

と表される．但し，$\alpha_{k,m}>0$ で，

$$\sum_{k=1}^{K}\sum_{m=1}^{M_k}\alpha_{k,m}=1 \tag{2}$$

$$g(x;\mu^{(k,m)},\Sigma^{(k,m)}) =(2\pi)^{-d/2}\,|\Sigma^{(k,m)}|^{-1/2} \times\exp\left[-\frac{1}{2}(x-\mu^{(k,m)})^T(\Sigma^{(k,m)})^{-1}(x-\mu^{(k,m)})\right] \tag{3}$$

である．但し，$|\cdot|$ は行列式を表す．

例えば，特徴ベクトル x を K 個のクラスに識別する問題を考えよう．ここで，各事象は従属とし，同時に複数の事象が生じることはないとする．このとき，識別は事後確率の最大な事象を選べばよい．特徴ベクトル x が入力されたとき事象 k の事後確率 $P(k|x)$，$k=1,\cdots,K$ は，特徴ベクトル x の生起確率密度関数に混合

正規分布を用いれば，

$$P(k\mid x)=\sum_{k=1}^{K}P(k,m\mid x)$$

$$=\sum_{m=1}^{M_k}\frac{P(k,m)P(x\mid k,m)}{P(x)} \qquad (4)$$

となる．ここで，M_k は事象 k のコンポーネント数，$P(k,m)$ は事象 k，コンポーネント m の事前確率で混合度 $\alpha_{k,m}$ に対応する．また，$P(x\mid k,m)$ は事象 k，コンポーネント m によって条件づけられた x の生起確率である．このとき事後確率 $P(k,m\mid x)$ は，式(1)を用いて

$$P(k,m\mid x)=\frac{P(k,m)P(x\mid k,m)}{\sum_{k'=1}^{K}\sum_{m'=1}^{M_{k'}}P(k',m')P(x\mid k',m')}$$

$$=\frac{\alpha_{k,m}g(x;\mu^{(k,m)},\Sigma^{(k,m)})}{\sum_{k'=1}^{K}\sum_{m'=1}^{M_{k'}}\alpha_{k',m'}g(x;\mu^{(k',m')},\Sigma^{(k',m')})} \qquad (5)$$

のように表現できる．

2.2 対数線形モデルの導入

式(5)の右辺の分子 $\alpha_{k,m}g(x;\mu^{(k,m)},\Sigma^{(k,m)})$ は，$g(x;\mu^{(k,m)},\Sigma^{(k,m)})$ が d 次元正規分布であるから(式(3))，平均ベクトル $\mu^{(k,m)}=[\mu_1{}^{(k,m)},\cdots,\mu_d{}^{(k,m)}]^T$，共分散逆行列 $\Sigma^{(k,m)-1}=[s_{ij}{}^{(k,m)}]$ を用いて

$$\alpha_{k,m}g(x;\mu^{(k,m)},\Sigma^{(k,m)})$$

$$=\exp\Big[-\frac{1}{2}\sum_{j=1}^{d}\sum_{l=1}^{j}(2-\delta_{jl})s_{jl}{}^{(k,m)}x_jx_l$$

$$+\sum_{j=1}^{d}\sum_{l=1}^{d}s_{jl}{}^{(k,m)}\mu_j{}^{(k,m)}x_l$$

$$-\frac{1}{2}\sum_{j=1}^{d}\sum_{l=1}^{d}s_{jl}{}^{(k,m)}\mu_j{}^{(k,m)}\mu_l{}^{(k,m)}-\frac{d}{2}\log 2\pi$$

$$-\frac{1}{2}\log\mid\Sigma^{(k,m)}\mid+\log\alpha_{k,m}\Big] \qquad (6)$$

と展開できる．但し，δ_{ij} は，クロネッカーのデルタで $i=j$ のとき 1，$i\neq j$ のとき 0 の値をとる．

ここで，式(6)の右辺を見掛け上，線形化することを考えよう．両辺の対数をとり，それを $\xi_{k,m}$ とおくと，

$$\xi_{k,m}\triangleq\log\alpha_{k,m}g(x;\mu^{(k,m)},\Sigma^{(k,m)})=\beta^{(k,m)T}X \qquad (7)$$

となる．但し，$X\in\mathcal{R}^H$，$\beta\in\mathcal{R}^H$ は，

$$X=[1,x^T,x_1{}^2,x_1x_2,\cdots,x_1x_d,x_2{}^2,x_2x_3,$$
$$\cdots,x_2x_d,\cdots,x_d{}^2]^T \qquad (8)$$

$$\beta^{(k,m)}=\Big[\beta_0{}^{(k,m)},\sum_{j=1}^{d}s_{j1}{}^{(k,m)}\mu_j{}^{(k,m)},\cdots,$$

$$\sum_{j=1}^{d}s_{jd}{}^{(k,m)}\mu_j{}^{(k,m)},-\frac{1}{2}s_{11}{}^{(k,m)},-s_{12}{}^{(k,m)},\cdots,$$

$$-s_{1d}{}^{(k,m)},\cdots,-\frac{1}{2}(2-\delta_{jl})s_{jl}{}^{(k,m)},$$

$$\cdots,-s_{d1}{}^{(k,m)},\cdots,-\frac{1}{2}s_{dd}{}^{(k,m)}\Big]^T \qquad (9)$$

で，$\beta_0{}^{(k,m)}=-\frac{1}{2}\sum_{j=1}^{d}\sum_{l=1}^{d}s_{jl}{}^{(k,m)}\mu_j{}^{(k,m)}\mu_l{}^{(k,m)}-\frac{d}{2}\log 2\pi-\frac{1}{2}\log\mid\Sigma^{(k,m)}\mid+\log\alpha_{k,m}$ で与えられる．いずれも $H=1+d(d+3)/2$ 個の要素からなるベクトルである．すなわち $\xi_{k,m}$ を係数ベクトル $\beta^{(k,m)}$ と新しく定義した入力ベクトル $X\in\mathcal{R}^H$ の積の形で表現できることがわかる．

しかしながら，コンポーネントの事後確率 $P(k,m\mid x)$ の総和 $\sum_{k=1}^{K}\sum_{m=1}^{M_k}P(k,m\mid x)$ は 1 であるから，事後確率のパラメータとして $\xi_{k,m}$ は冗長である．そこで，新しく変数 $Y_{k,m}$ と係数ベクトル $w^{(k,m)}\in\mathcal{R}^H$ を導入し，

$$Y_{k,m}\triangleq\xi_{k,m}-\xi_{K,M_K}=(\beta^{(k,m)}-\beta^{(K,M_K)})^TX=w^{(k,m)T}X \qquad (10)$$

とおく．但し，定義から $w^{(K,M_K)}=0$ である．これにより，$w^{(k,m)}$ を無制約の重み係数とみなすことが可能となる．このとき式(5)の事後確率 $P(k,m\mid x)$ は

$$P(k,m\mid x)=\frac{\exp[Y_{k,m}]}{\sum_{k'=1}^{K}\sum_{m'=1}^{M_{k'}}\exp[Y_{k',m'}]} \qquad (11)$$

となる．

以上のように，各コンポーネントの確率密度関数の対数をとることで，事後確率を式(10)，式(11)のように入力ベクトル X と係数ベクトル $w^{(k,m)}$ の線形和である変数 $Y_{k,m}$ を用いて表現することができた．これは，混合正規分布に対数線形モデル[10]を導入したことに相当し，これにより混合正規分布モデルに含まれるパラメータを，より少ない数のパラメータ $w^{(k,m)}$ に置き換えることが可能となった．

以上の定式化は，共分散行列が特殊な構造をもつ場合にも適用可能である．例えば共分散行列 $\Sigma^{(k,m)}\in\mathcal{R}^{d\times d}$ が単位行列の定数倍で，事象 k，コンポーネント m によらずこの定数が等しい場合を考えよう．この場合式(10)は，共分散行列の定数を s^{-1} とすると

$$Y_{k,m}=s\sum_{j=1}^{d}\Big(\mu_j{}^{(k,m)}-\mu_j{}^{(K,M_K)}x_j$$

$$-\frac{1}{2}s\sum_{j=1}^{d}(\mu_j{}^{(k,m)2}-\mu_j{}^{(K,M_K)2})+\log\frac{\alpha_{k,m}}{\alpha_{K,M_K}}$$

$$= w^{(k,m)T} X \tag{12}$$

となる．但し，入力ベクトル X と重みベクトル $w^{(k.,m)}$ は，

$$X=[1, x^T]^T \in \mathcal{R}^{d+1} \tag{13}$$

$$w^{(k,m)}=\Big[-\frac{1}{2}s\sum_{j=1}^{d}(\mu_j{}^{(k,m)2}-\mu_j{}^{(K,M_K)2}) + \log\frac{\alpha_{k,m}}{\alpha_{K,M_K}}, s(\mu_1{}^{(k,m)}-\mu_1{}^{(K,M_K)}), \cdots, s(\mu_d{}^{(k,m)}-\mu_d{}^{(K,M_K)})\Big]^T \in \mathcal{R}^{d+1} \tag{14}$$

である．このように，対象にする分布の特性に対応して，入力ベクトル X と重みベクトル $w^{(k,m)}$ の構造が自然に決定されることがわかる．次章では，この定式化をフィードフォワード型ニューラルネットに展開する．

3. フィードフォワード型ニューラルネットによる事後確率の推定

3.1 ニューラルネットの構造

本研究で提案するニューラルネットの構造図を図1に示す．まず，特徴ベクトル $x \in \mathcal{R}^d$ を前処理しニューラルネットへの入力ベクトル $X \in \mathcal{R}^H$ に変換する（式(8)）．ニューラルネットの第1層はこの入力ベクトル X の次元数 H に合わせて H 個のユニットからなり，ユニットの入出力関数には恒等関数を用いる．第1層の入出力関係は，入力を ${}^{(1)}I_j$，出力を ${}^{(1)}O_j$ とすれば，

$${}^{(1)}I_j = X_j \tag{15}$$

$${}^{(1)}O_j = {}^{(1)}I_j \tag{16}$$

となる．

第2層は混合正規分布の総コンポーネント数 $\sum_{k=1}^{K} M_k$ と同数のユニットからなり，第1層の出力を重み係数 $w_h{}^{(k,m)}$ を介して受け取り，式(11)に従って各コンポーネントの事後確率を出力する．第2層のユニット $\{k,m\}$ への入力を $Y_{k,m}$ とし，出力を ${}^{(2)}O_{k,m}$ とすれば，

$$Y_{k,m} = \sum_{h=1}^{H} {}^{(1)}O_h w_h{}^{(k,m)} \tag{17}$$

$${}^{(2)}O_{k,m} = \frac{\exp[Y_{k,m}]}{\sum_{k'=1}^{K}\sum_{m'=1}^{M_K}\exp[Y_{k',m'}]} \tag{18}$$

となる．但し，$w^{(K,M_K)}=0$ である．

第3層は，事象数 K 個のユニットからなり，事象 k $(k=1,\cdots,K)$ の事後確率を出力する．ユニット k は，第2層の M_k 個のユニット $\{k,m\}$ $(m=1,\cdots,M_K)$ と結合している．入出力関係は，

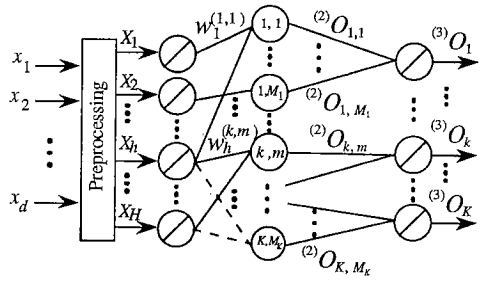

図1 本ニューラルネットの構造
Fig. 1 Structure of the proposed neural network.

$${}^{(3)}I_k = \sum_{m=1}^{M_k} {}^{(2)}O_{k,m} \tag{19}$$

$${}^{(3)}O_k = {}^{(3)}I_k \tag{20}$$

である．

以上，本論文で提案するニューラルネットの構造について説明した．このニューラルネットでは，第1層と第2層の間の結合重み係数 $w_h{}^{(k,m)}$ を学習的に推定するだけで，各事象の事後確率を混合正規分布モデルに基づいて計算できることになる．

3.2 ネットワークの学習則

今，n 番目特徴ベクトル $x^{(n)}$ に対して，観測ベクトル $T^{(n)}=[T_1{}^{(n)},\cdots,T_k{}^{(n)},\cdots,T_K{}^{(n)}]^T$ が与えられた場合を考えよう．$T_k{}^{(n)}$ は，観測された事象が k であるときに1，それ以外のときは0をとり，他の事象と同時に1にはならない．ある特徴ベクトル $x^{(n)}$ が入力されたとき観測ベクトル $T^{(n)}$ が得られる確率は，事後確率 $P(k \mid x^{(n)})$ を用いて，

$$P(T^{(n)}) = \prod_{k=1}^{K} P(k \mid x^{(n)})^{T_k{}^{(n)}} \tag{21}$$

で与えられる．

ネットワークは，N 個のサンプルデータ $x^{(n)}$ $(n=1,\cdots,N)$ を用いて学習を行う．N 個のサンプルデータ（学習用データ）が与えられたときの対数ゆう度関数 L は，式(21)を用いると，

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} T_k{}^{(n)} \log {}^{(3)}O_k \tag{22}$$

となる．ニューラルネットの出力値 ${}^{(3)}O_k$ は事後確率 $P(k \mid x^{(n)})$ に対応する．評価関数 J としては，式(22)にマイナスを付けた．

$$J = \sum_{n=1}^{N} J_n = -\sum_{n=1}^{N}\sum_{k=1}^{K} T_k{}^{(n)} \log {}^{(3)}O_k \tag{23}$$

を用い，これを最小化，すなわちゆう度を最大化するように学習を行う．このとき，特徴ベクトル $x^{(n)}$ が入

力されたときの重み $w_h^{(k,m)}$ $(h=1,\cdots,H)$ の修正量 $\Delta w_h^{(k,m)}$ は，η を学習率とすると，逐次修正型の場合，

$$\Delta w_h^{(k,m)}=-\eta\frac{\partial J_n}{\partial w_h^{(k,m)}} \tag{24}$$

一括修正型の場合，

$$\Delta w_h^{(k,m)}=-\eta\sum_{n=1}^{N}\frac{\partial J_n}{\partial w_h^{(k,m)}} \tag{25}$$

となる．但し，

$$\eta\frac{\partial J_n}{\partial w_h^{(k,m)}}=\frac{\partial}{\partial w_h^{(k,m)}}\left(-\sum_{k=1}^{K}T_k^{(n)}\log{}^{(3)}O_k\right)$$
$$=\left({}^{(2)}O_{k,m}-\frac{{}^{(2)}O_{k,m}}{{}^{(3)}O_k}T_k^{(n)}\right)X_h^{(n)} \tag{26}$$

である．また各コンポーネントに教師信号が与えられた場合には，

$$\frac{\partial J_n}{\partial w_h^{(k,m)}}=\frac{\partial}{\partial w_h^{(k,m)}}\left(-\sum_{k=1}^{K}\sum_{m=1}^{M_K}T_{k,m}^{(n)}\log{}^{(2)}O_{k,m}\right)$$
$$=({}^{(2)}O_{k,m}-T_{k,m}^{(n)})X_h^{(n)} \tag{27}$$

としてこう配方向の計算をすればよい．但し，$T_{k,m}$ はコンポーネント$\{k,m\}$に対する教師信号で，観測された事象がコンポーネント$\{k,m\}$に対応するとき1，それ以外のとき0をとり，他のコンポーネントと同時に1にはならない．一般にパターン識別問題では，事象に対して教師信号が与えられる場合が多く，各コンポーネントに対する教師信号が与えられることはまれである．学習則式(26)は，事象に対して与えられた教師信号を，各コンポーネントの事後確率 ${}^{(2)}O_{k,m}$ が事象 k の事後確率 ${}^{(3)}O_k$ に占める割合に応じて分配した形になっていることがわかる．

本論文で示した学習則は，教師信号 $T_k^{(n)}$ が$\{0,1\}$の離散値である場合だけでなく，$[0,1]$の連続値の場合にもそのまま適用することができる(但し，$T_k^{(n)}\geqq 0$ で $\sum_{k=1}^{K}T_k^{(n)}=1$ とする)．今，教師信号ベクトル $T^{(n)}$ とニューラルネットからの出力ベクトル ${}^{(3)}O=[{}^{(3)}O_1,\cdots,{}^{(3)}O_k,\cdots,{}^{(3)}O_K]^T$ が与えられたとき，Kullbackの情報量[12]を用いて学習のための評価関数を

$$J'=\sum_{n=1}^{N}I(T^{(n)};{}^{(3)}O)$$
$$=\sum_{n=1}^{N}\sum_{k=1}^{K}T_k^{(n)}\log\frac{T_k^{(n)}}{{}^{(3)}O_k}$$
$$=\sum_{n=1}^{N}\sum_{k=1}^{K}T_k^{(n)}\log T_k^{(n)}$$
$$-\sum_{n=1}^{N}\sum_{k=1}^{K}T_k^{(n)}\log{}^{(3)}O_k\geqq 0 \tag{28}$$

のように定義する．上式の第1項は定数となるので，${}^{(3)}O$ を $T^{(n)}$ に近づけるには，第2項を最小化すればよい．この第2項が式(23)の J と同じであることに注意すると，式(24)～(27)がKullbackの情報量を最小にするための学習則を与えていることがわかる．

4. 評価実験

4.1 一般化能力——誤差逆伝搬型ニューラルネットとの比較——

本論文で提案したニューラルネットと誤差逆伝搬型ニューラルネットとの一般化能力を比較するため，人工的に作成したデータを用いて識別実験を行った．用いたデータは2事象 $(K=2)$ で，各事象とも二つのコンポーネントをもつ混合正規分布を用いて作成した $(M_1=M_2=2)$．但し，特徴ベクトル x の次元数は2である $(d=2,H=6)$．用いた混合正規分布のパラメータを表1に示す．

本ニューラルネットは，データの総コンポーネント数にあわせて第2層に4個のユニットを用意し，入力層のユニットは6個，出力層は2個とした．一方，誤差逆伝搬型ニューラルネットは，特徴ベクトル x の次元数にあわせて入力層には線形ユニットを2個，中間層(2層)にはそれぞれ10個，出力層には2個のシグモイド関数ユニットを用意した．また二つのニューラル

表1 混合正規分布のパラメータ
Table 1 Parameters of the Gaussian mixture model used in the experiments.

		$\alpha_{k,m}$		$\mu^{(k,m)T}$		$\Sigma^{(k,m)}$	
component, m		1	2	1	2	1	2
event, k	1	0.3	0.2	[0.0,0.0]	[5.0,7.0]	9.0 0.63 0.63 0.09	1.0 -0.5 -0.5 1.0
	2	0.25	0.25	[2.0,2.0]	[−6.0,2.0]	1.0 2.7 2.7 9.0	0.09 0.0 0.0 1.0

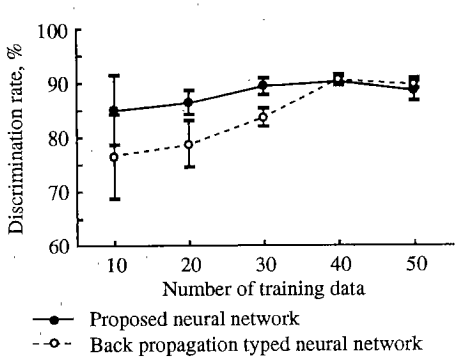

図2 学習用データ数による識別率の変化
Fig. 2 Effect of the number of training data on discrimination ability.

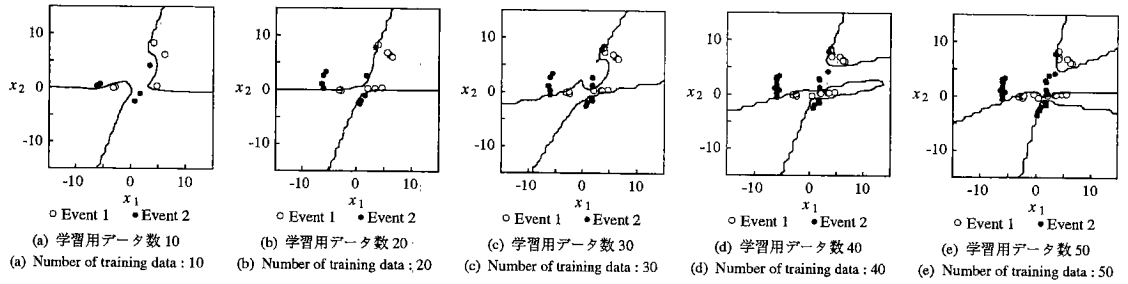

図3 学習用データと本手法による識別曲線
Fig. 3 Scatter diagram of training data and decision region boundaries learned in the proposed network.

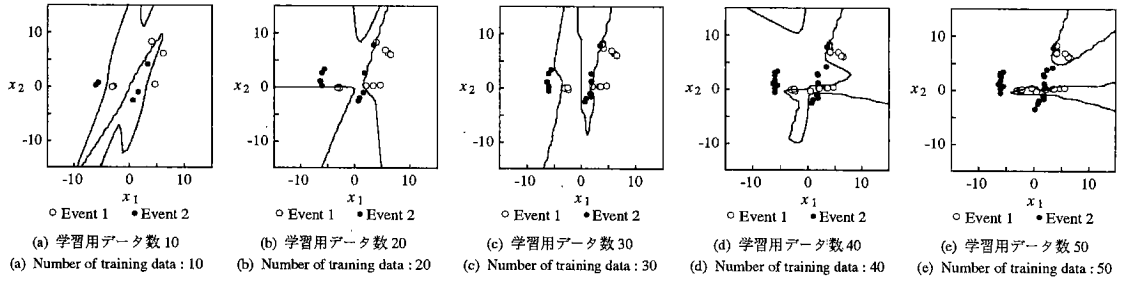

図4 学習用データと誤差逆伝搬型ニューラルネットの識別曲線
Fig. 4 Scatter diagram of training data and decision region boundaries learned in the back propagation network.

ネットに同等の学習を行わせるため，すべての学習用データに対して式(23)の評価関数 J_n の値が，0.1 以下になるまで学習を行わせた．但し，教師信号は各事象に対して与え (式(24)，式(26)；$\eta=0.0001$)，誤差逆伝搬型ニューラルネットは出力の和が 1 になるように正規化した．

図 2 に，学習用データの数による識別率の変化を示す．図は，2 種類のニューラルネットを 10 個から 50 個までの 5 通りの学習用データを用いて学習し，それとは別に作成した各事象 1000 個，計 2000 個のデータに対してネットワークが正しく識別した割合で，いずれも 10 種類の初期重みに対する平均と標準偏差を示している．実線 (黒丸) が本手法，点線 (白丸) が誤差逆伝搬型ニューラルネットでの結果である．図から，学習用データ数が多いときにはどちらのネットワークも高い識別率を示しているが，学習用データの数が減少するにつれて二つのネットワークの識別率に差が生じているのがわかる．誤差逆伝搬型ニューラルネットでは学習用データ数の減少に伴い識別率が大きく低下しているが，本ネットワークでは少ない学習用データでも高い識別率を示している．

このときの学習用データの散布図とニューラルネットで学習した事象の識別曲線 (各事象の事後確率が等しくなる境界) を図 3 (本ネットワーク)，図 4 (誤差逆伝搬型ニューラルネット) に示す．誤差逆伝搬型ニューラルネットでは識別曲線が学習用データ数によって大きく変化しているのに対し，本ネットワークではほぼ同じ形の識別曲線が学習によって獲得できている．

4.2 表現能力——混合正規分布モデルとの比較——

混合正規分布モデルのパラメータ (混合度 $\alpha_{k,m}$，平均ベクトル $\mu^{(k,m)}$，共分散行列 $\Sigma^{(k,m)}$) は，その性質からとり得る値に制約を受ける．例えば，共分散行列の要素は逆行列が存在する値でなければならないし，混合度 $\alpha_{k,m}$ は正でかつ総和が 1 でなければならない．一方，本手法のパラメータである重み係数 $w_h^{(k,m)}$ は無制約で互いに独立である．そこでこの違いを評価するため，混合正規分布モデルの理論をほぼそのままニューラルネットに展開した MLANS (Maximum Likelihood Artificial Neural System)[6] と本ネットワークとの比較実験を行った．

識別に用いたデータは人工的に作成した 3 事象のデータで ($K=3$)，事象 A，B が一様分布に，事象 C が二つの一様分布の混合分布に従う．その学習用データの一例を図 5 に示す．特徴ベクトル x の次元数は 2 であ

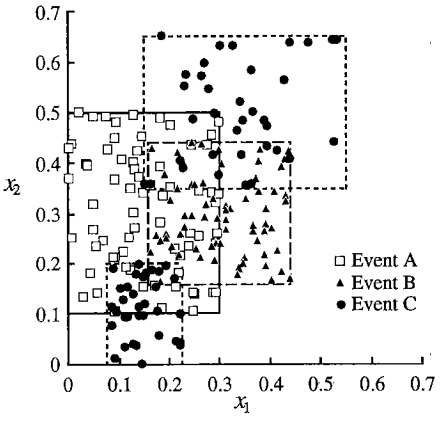

図5　学習用データの散布図(210個)
Fig. 5　Scatter diagram of 210 training data.

る($d=2, H=6$)．本ニューラルネットは入力層のユニットは6個，出力層のユニットは事象数に合わせて3個，第2層のユニットはMLANSの総コンポーネント数に合わせて用意した．教師信号は各事象に与え(式(24)，式(26)：$\eta=0.5$)，式(23)の評価関数Jの学習用データ数に対する平均値が0.5以下になるまで学習を行った．なお，学習用データ数が210，270，330個については，学習の高速化を行うため，Wangら[13]の方法を参考にし本学習則にTerminal Attractorを適用して学習を行った．一方，MLANSは繰返し学習に伴う事後確率の変化がBhattacharyya距離[12]で0.0001以下になるまで学習を行わせた．識別に用いたデータは，学習用データとは別に作成した各事象1000個，計3000個のデータである．

図6は学習用データ数を30個から330個まで変化させた場合の識別結果である．それぞれ10種類の初期重みに対する平均と標準偏差を示す．実線(黒丸)が本手法，点線(白丸)がMLANSでの結果である．なお，MLANSの総コンポーネント数および本ニューラルネットの第2層のユニット数は，9個(各事象3個)としている．学習用データ数が多い場合はどちらの方法ともうまく識別できているが，学習用データ数が減少するにつれて本手法に比べてMLANSの識別率が低下していることがわかる．なお，学習用データ数が30個の場合は，MLANSに含まれる共分散行列を計算することができなかった．これは，学習用データ数が少ないとコンポーネントに含まれるデータが非常に少なくなる場合が生じるためである．

一方，図7はコンポーネント数を各事象同数にし，

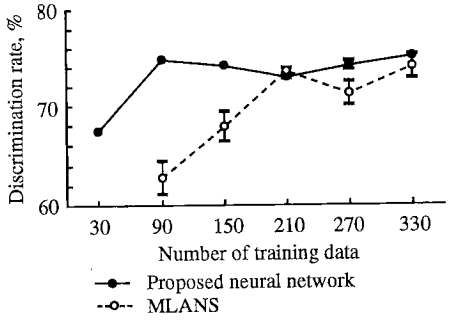

図6　学習用データ数による識別率の変化
Fig. 6　Effect of the number of training data on discrimination ability.

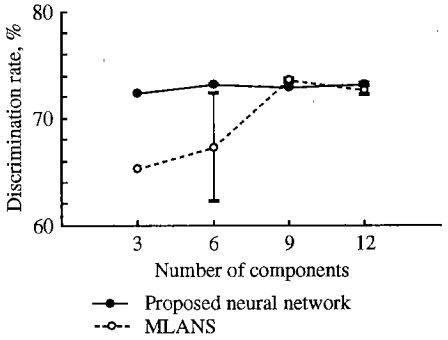

図7　コンポーネント数による識別率の変化
Fig. 7　Effect of the unmber of components on discrimination ability.

総コンポーネント数を3個から12個まで変化させた場合の識別結果である．学習用データは210個(各事象70個)とし，図6と同じ学習終了判定を用いて学習を行った．但し，コンポーネント数3の場合は，本手法では評価関数Jの学習用データ数に対する平均値を0.5以下になるまで学習することができなかったので，0.6以下まで学習した結果を示した．MLANS(点線，白丸)では，十分コンポーネント数が多い場合にはうまく識別できているが，コンポーネント数が少なくなるとデータの分布を適切に表現できなくなり識別率が低下している．図中，コンポーネント数が6個のとき標準偏差が大きくなっているのは，初期重みによって学習結果が三つに分かれてしまったためである．一方，本手法では，MLANSと比べて少ないコンポーネント数でも識別がうまく行えている．

5. むすび

本論文では，フィードフォワード型ニューラルネッ

トの一般化能力を改善することを目的として，対数線形モデルを用いて混合正規分布の事後確率を推定する新しいニューラルネットを提案した．まず，混合正規分布モデルに対数線形モデルを導入することで，より少ないパラメータで事後確率を計算できることを示した．次にこの前向き計算をニューラルネットに展開し，ゆう度最大の評価のもとで学習則を導出した．本ネットワークは，

（1） 通常の誤差逆伝搬型ニューラルネットに入力データの分布を近似する統計構造を組み込んだ形になっているので，少ない学習用データや学習用データとして与えられていない入力データに対しても高い識別能力を実現できる，

（2） 混合正規分布モデルのパラメータを無制約化することで，より忠実に入力データの特性を表現できる，

（3） 統計手法を組み込むことで，ユニットの入出力関数，層の数，ユニットの数などのネットワーク構造を自然に決定することができる，

（4） ニューラルネットからの出力値を確率として取り扱うことができる

などの特徴を有している．

今後は，学習過程に応じてコンポーネント数を適応的に調節する方法，学習の高速化，および本ネットワークの実データへの適用について考察する予定である．

文　献

（1） Rumelhart D. E., McClelland J. L. and Williams R. J. : "Learning Internal Representations by Error propagation", in Parallel Distributed Processing **I**, pp. 318-362, MIT Press (1986).

（2） Bridle J. S. : "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships of Statistical Pattern", Neurocomputing : Algorithms, Architectures and Applications, Suilie F. F. and Herault J., pp. 227-236, Springer-Varlag (1990).

（3） 中川聖一，小野義之："ニューラルネットワークによる確率密度関数・事後確率の推定と母音認識"，信学論(D-II), **J76-D-II**, 6, pp. 1081-1089 (1993-06).

（4） Traven H. G. C. : "A Neural Network Approach to Statistical Pattern Classification by Semiparametric Estimation of Probability Density Functions", IEEE Trans. Neural Networks, **2**, 3, pp. 366-377 (May 1991).

（5） Everitt B. S. and Hand D. J. : "Finite Mixture Distributions", Chapman and Hall (1981).

（6） Perlovsky L. I. and McManus M. M. : "Maximum Likelihood Neural Networks for Sensor Fusion and Adaptive Classification", Neural Networks, **4**, pp. 89-102 (1991).

（7） 辻　敏夫，森大一郎，伊藤宏司："統計的構造を組込んだニューラルネットによるEMG動作識別法"，電学論，**112-C**, 8, pp. 465-473 (1992-08).

（8） Lee S. and Shimoji S. : "Self-Organization of Gaussian Mixture Model for Learning Class pdfs in Pattern Classification", Proc. IJCNN, **3**, pp. 2492-2495 (1993).

（9） Jordan M. I. and Jacobs R. A. : "Hierarchical Mixtures of Experts and the EM algorithm", Proc. IJCNN, **2**, pp. 1339-1344 (1993).

(10) McCullagh P. and Nelder J. A. : "Generalized linear models", Chapman and Hall (1983).

(11) Dempster A. P., Laird N. M. and Rubin D. B. : "Maximum Likelihood from Incomplete Data via the EM Algorithm", J. Roy. Stat. Soc., **39**, 1, pp. 1-38 (1977).

(12) 竹内　啓，柳井晴夫："多変量解析の基礎"，東洋経済新聞社 (1972).

(13) Wang S. D. and Hsu C. H. : "Terminal Attractor Learning Algorithms for Back Propagation Neural Networks", Proc. IJCNN, **1**, pp. 183-189 (1991).

(平成6年3月9日受付)

辻　敏夫

昭60広島大大学院工学研究科博士課程前期了，同年同大学工学部助手．平4イタリア国ジェノバ大学客員研究員．人間とロボットの運動制御，ニューラルネット，マン・マシンシステムなどの研究に従事．電気学会，日本ロボット学会，IEEE等各会員（工博）．

市延　弘行

平4広島大・工・第二類（電気系）卒．現在，同大大学院工学研究科博士課程前期システム工学専攻在学中．ニューラルネットを用いたパターン識別に関する研究に従事

金子　真

昭51九工大卒．昭56東大工学系研究科博士課程卒．工博．同年4月，通産省工業技術院機械技術研究所入所．平2年4月，九州工業大学情報工学部助教授．平5年10月広島大教授，現在に至る．ロボットハンドやモーションベーストアクティブセンシングなどの研究に興味をもつ．IEEE，計測自動制御学会，日本ロボット学会等各会員．