数理解析研究所講究録
第870巻 1994年 56-63

情報量を導入した ニューラルネットワークＢＰ学習法

Back-Propagation with Maximum Information

千葉大学
Chiba University

宇野達也　平田廣則
Tatsuya Uno　Hironori Hirata

1. はじめに

階層型ニューラルネットワークの代表的な学習法であるバックプロパゲーション（ＢＰ）学習法は研究が盛んに行われ様々な分野に応用されている。[1] この学習法では階層間の結合は全結合で分散的なものになり、学習問題がどのようにネットワーク構造に反映しているかわかりにくい。そこで石川らはネットワークの結合荷重に対してエントロピーを定義し、そのエントロピーを減らすことによってネットワーク構造を単純化させる学習法を提案している。[2] [3] 本文ではエントロピーより高度なネットワーク構造の指標である情報量を定義する。ネットワーク構造が単純になるほど情報量は大きくなり、情報量を大きくすることをＢＰ学習に導入することにより単純なネットワーク構造が得られる。構造を持つ学習問題に対して計算機実験を行った結果、学習問題に対応したネットワーク構造が本学習法により得られることがわかった。

また、ＢＰ学習法の問題点としてネットワークのユニット数の適切な値を決めることが困難なことがある。ユニット数が多すぎても少なすぎてもうまく学習が行われない。本文では情報量を導入した学習法によりネットワークを単純化し、学習後一括してユニット数を決定する方法を提案する。計算機実験を行った結果、ＸＯＲ問題に対して、適切なユニット数が得られた。

2. バックプロパゲーション（ＢＰ）学習法

2.1. 階層型ニューラルネットワーク

階層型ニューラルネットワークを図1に示す。ネットワークは入力層、複数の中間層、出力層からなる。入力層に入った信号はユニットを結ぶ重みw_{ij}を通して中間層に達する。同様に信号は中間層から出力層に達し、ネットワークの出力となる。このとき、ｉ番目のユニットの出力をo_i、ｊ番目のユニットからｉ番目のユニットへの重みをw_{ji}、ｉ番目のユニットのしきい値をθ_iとすると、

$$o_i = f(\sum w_{ji} o_j + \theta_i) \qquad (1)$$

となる。ここで，fはシグモイド関数を用いる。

$$f(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

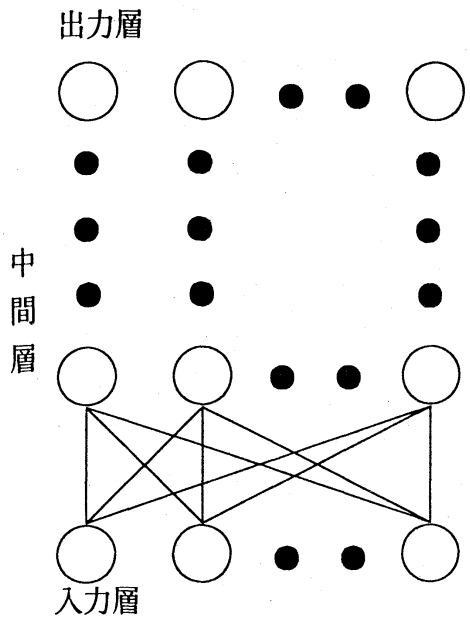

図 1 階層型ニューラルネットワーク

2.2. バックプロパゲーション学習法

学習問題として、入力層のユニットへの入力の値とそれに対する望ましい出力層のユニットの出力である教師信号の値のパターンをネットワークに提示する。ネットワークは、入力層に与えられた入力パターンを出力層に伝播し、出力パターンを出す。それを望ましい出力パターンと比較し、得られた出力と教師信号の差を減らすようにネットワークの結合の重みを調整することにより学習を行う。

その差を減らすため誤差の二乗和である二乗誤差Eを減らす。k番目の学習パターンのi番目の出力層のユニットに対する教師信号をT_{ki}、i番目の出力層のユニットの出力をo_iとすると、Eは

$$E = \frac{1}{2}\sum_{k=1}^{P}\sum_{i=1}^{NO}(T_{ki}-o_i) \qquad (3)$$

と表される。ここで、*P*はパターン数、NOは出力層のユニット数を表す。

この二乗誤差を減らす方法として最急降下法を用いる。重みw_{ij}の修正量をΔw_{ij}とすると、

$$\Delta w_{ij} = -\eta\frac{\partial E}{\partial w_{ij}} \qquad (4)$$

となる。

また、一般に学習を速くするため、一回前の重みの変化量を$\Delta w'_{ij}$とし、重みの変化量に慣性項と呼ばれる項を加える。

$$\Delta w_{ij} = -\eta\frac{\partial E}{\partial w_{ij}} + \alpha\Delta w_{ij}/ \qquad (5)$$

ここで、αは定数である。

3. 情報量を導入したＢＰ学習法

3.1. ネットワークにおける確率の定義

確率の定義は、入力層から中間層などのある層からある層へ信号がある時刻に同時に次のユニットへ重みの絶対値の大きさで伝わるとみなしている。この層間を通信路のようにみなし各層間ごとに確率を定義する。ユニットｉからユニットｊへの重みをw_{ij}、二つの層の多いほうのユニット数をＮとすると、ユニットｉが信号を出力する確率q_iを

$$q_i=\frac{\sum_{j=1}^{N}|w_{ij}|}{W} \tag{6}$$

$$W=\sum_{i=1}^{N}\sum_{j=1}^{N}|w_{ij}| \tag{7}$$

とする。また、ユニットｉが信号を受け取る確率p_iを

$$p_i=\frac{\sum_{j=1}^{N}|w_{ji}|}{W} \tag{8}$$

とする。ユニットｉからユニットｊへ信号が伝わる確率p_{ij}を

$$p_{ij}=\frac{|w_{ij}|}{\sum_{j=1}^{N}|w_{ij}|} \tag{9}$$

とする。

3.2. ネットワークの情報量の定義

(1) 等確率性からのずれD_1

これは、ユニットごとに情報がどのくらい偏って伝達されているかを表す。D_1をユニットが信号を受け取る確率がすべてのユニットに対して等確率なときつまり、最大のエントロピーH_{1max}からそのネットワークのエントロピーH_1を引いたものと定義する。

$$D_1=H_{1max}-H_1 \tag{10}$$

$$H_1=-\sum_{i=1}^{N}p_i\log p_i \tag{11}$$

$$H_{1max}=\log N \tag{12}$$

D_1は図 2に示すようにすべてのユニットに対して、信号の流れが分散しているとき小さく、図 3に示すようにあるユニットに対して信号の流れが集中しているとき大きくなる。情報量D_1を増やすことによりネットワークの動作に必要なユニット数を少なくすることができる。

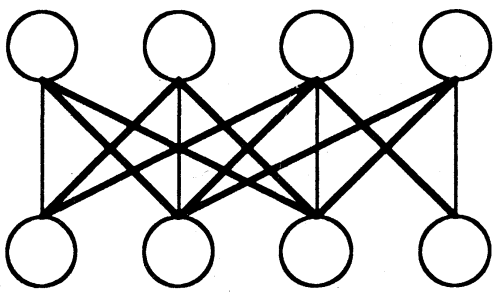

図 2 D_1が小さいネットワーク

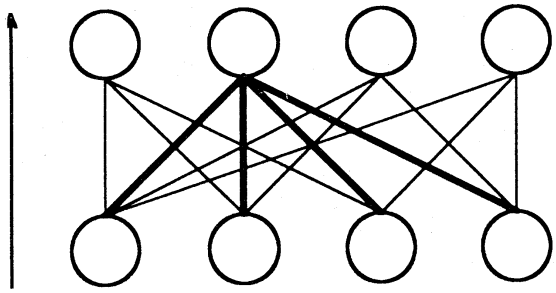

図 3　D_1が大きいネットワーク

(2)　独立性からのずれD_2

これは、信号を受け取るユニットから見てどのくらい信号の流れが偏って信号を受け取っているかを表すものである。あるユニットからあるユニットに対して信号が伝わる結合確率のエントロピーを考え、D_2をその確率が独立なときのエントロピーH^{IND}からネットワークのエントロピーH^Dを引いたものと定義する。

$$\begin{aligned} D_2 &= H^{IND} - H^D \\ &= -\sum_{i=1}^{N}\sum_{j=1}^{N} q_i\, p_j \log q_i\, p_i \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} q_i\, p_{ij} \log q_i\, p_{ij} \\ &= H_1 - H_M \end{aligned} \tag{13}$$

$$H_M = -\sum_{i=1}^{N}\sum_{j=1}^{N} q_i\, p_{ij} \log p_{ij} \tag{14}$$

D_2は図 4に示すように信号を受け取るユニットが等確率で信号を受け取るとき小さくなる。逆に、図 5に示すように結合が単純になるとD_2は大きくなる。このように、D_2を大きくすると単純なネットワーク構造を作ることができる。

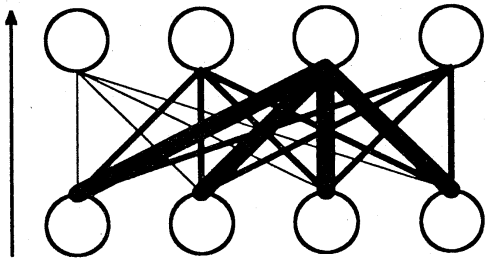

図 4　D_2が小さいネットワーク

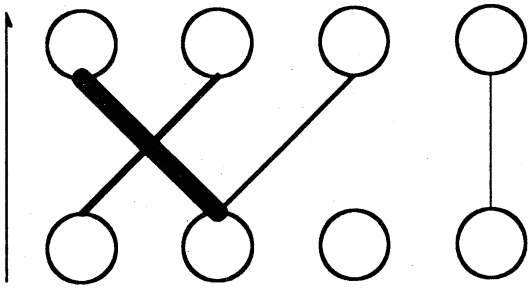

図 5　D_2が大きいネットワーク

(3)　全情報量ＴＩ

ＴＩをこれまで定義した二つの情報量D_1、D_2を用いて次のように定義する。

$$\begin{aligned} TI &= D_1 + D_2 \\ &= H_{1\max} - H_M \end{aligned} \tag{15}$$

このＴＩはD_1とD_2の両方のネットワークの特徴を表す。これら３つの情報量の関係を図 6に示す。ＴＩを増やすことで動作に必要なユニット数を少なくし、ネットワークの結合を単純化できる。

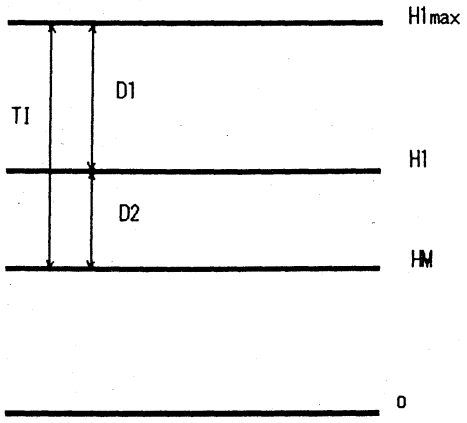

図 6 各情報量の関係

3.3. ＢＰ学習法への各情報量の導入

ＢＰ学習では、二乗誤差を減らすことにより学習を行う。しかし、情報量は大きくなるとネットワーク構造が単純になるため、情報量を増加するようにＢＰ学習に導入する。　用いる情報量はD_1とＴＩとする。

3.3.1.情報量D_1の導入

最小化すべき評価関数Ｊを

$$J = E - \lambda' D_1 \tag{16}$$

とする。

これをＢＰ学習と同様に最急降下法を用いて、重みを変更しＪを減らす。重みの変化量Δw_{ij}は

$$\begin{aligned}\Delta w_{ij} &= -\eta \frac{\partial J}{\partial w_{ij}} \\ &= -\eta \frac{\partial E}{\partial w_{ij}} + \eta \lambda' \frac{\partial D_1}{\partial w_{ij}} \\ &= -\eta \frac{\partial E}{\partial w_{ij}} - \lambda \frac{\partial H_1}{\partial w_{ij}} \qquad (\lambda = \eta \lambda')\end{aligned} \tag{17}$$

となる。ここで、λ'，λは情報量を導入する割合を表す。

また、このままD_1を増やすとある一つのユニットにつながる重みが限りなく大きくなってしまう。そうなると、本来の目的である二乗誤差が減らなくなり学習がうまく行われなくなったり、ネットワークの構造の解析においてもよい結果が得られない。これを防ぐためにD_1の微分値の符号と重みの符号が違う場合は$\lambda = 0$としてD_1の微分値を加えない。つまり、情報量の項に関しては重みの絶対値を増やさないものとする。さらに、ＢＰ学習法と同様に慣性項を加える。

3.3.2. 情報量ＴＩの導入

D_1と同様に評価関数Ｊを

$$J = E - \lambda' TI \tag{18}$$

とする。Δw_{ij}は

$$\begin{aligned}\Delta w_{ij} &= -\eta \frac{\partial J}{\partial w_{ij}} \\ &= -\eta \frac{\partial E}{\partial w_{ij}} + \eta \lambda' \frac{\partial TI}{\partial w_{ij}} \\ &= -\eta \frac{\partial E}{\partial w_{ij}} - \lambda \frac{\partial H_M}{\partial w_{ij}} \qquad (\lambda = \eta \lambda')\end{aligned} \tag{19}$$

となる。D_1と同様に微分値の符号と重みの符号が違う場合は$\lambda = 0$として微分値を加えない。さらに、ＢＰ学習法と同様に慣性項を加える。

3.3.3. エントロピー最小化学習法

エントロビー最小化学習法は各層間ごとにエントロピーを定義する。

$$p_{ij} = \frac{|w_{ij}|}{W} \quad (20)$$

$$W = \sum_i \sum_j |w_{ij}| \quad (21)$$

$$H = -\sum_i p_{ij} \log p_{ij} \quad (22)$$

このエントロピーを最急降下法を用いて少なくすることをＢＰ学習に加えた手法である。

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} - \lambda \frac{\partial H}{\partial w_{ij}} + \alpha \Delta w_{ij} / \quad (23)$$

αとλは定数である。また、重みの絶対値が増える場合エントロピーの微分値の項を加えない。さらに、ＢＰ学習法と同様に慣性項を加える。

4. 計算機実験

4.1. D_1の最大化

入力層と中間層の間のD_1を増やすことによって中間層のユニット数を決定する。学習問題をＸＯＲ、入力層のユニット数を２、中間層のユニット数を１０、出力層のユニット数を１、$\eta = 0.6$、$\lambda = 0.01$、$\alpha = 0.5$、学習の終了条件を二乗誤差Ｅが０．０５より小さくなったときとする。この条件のもとで実験を行った。

ＢＰ学習の結果を図 7に情報量を導入した学習の結果を図 8に示す。ＢＰ学習は分散的な構造になっているが、情報量を導入した学習では構造が単純になり二つのユニットに集中して重みが集まり他のユニットへは重みがなくなっている。このような重みがなくなった中間層のユニットを削除することにより中間層のユニット数を決定する。そこで、学習後入力層からの重みの絶対値の総和が０．１より小さい中間層のユニットを削除する。この削除を行う１００回の実験において中間層のユニット数の平均値は２．５８個となった。ＸＯＲに対する最適値は２個であるので、本学習法の有効性が確かめられたといえよう。

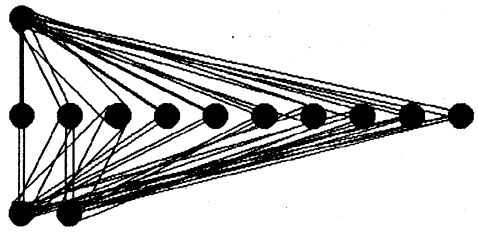

図 7 ＢＰ学習によるネットワーク

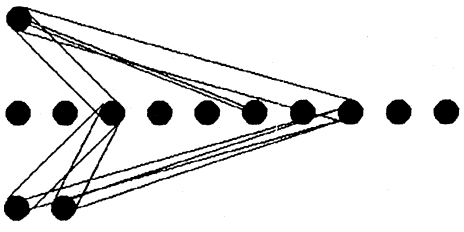

図 8 情報量によるネットワーク

4.2. ＴＩの最大化

入力層と中間層の間、中間層と出力層の間のＴＩを増やすことによって学習問題に対応したネットワーク構造を作る。
学習問題を式(24)に示す論理関数を用いる。

$$f = (a \cup c) \cap (b \cup d) \qquad (24)$$

入力層のユニット数を４、中間層のユニット数を４、出力層のユニット数を１、$\eta = 0.6$、$\lambda = 0.01$、$\alpha = 0.5$、学習の終了条件を二乗誤差Ｅが０．０５より小さくなったときとする。この条件のもとで実験を行った結果をＢＰ学習を図 9にエントロピー最小化学習を図 11に、情報量を導入した学習を図 10に示す。ＢＰ学習は分散的な構造になっているが、エントロピー最小化学習と情報量を導入した学習では学習問題に対応した構造ａとｃ、ｂとｄの関係を表す構造になっている。

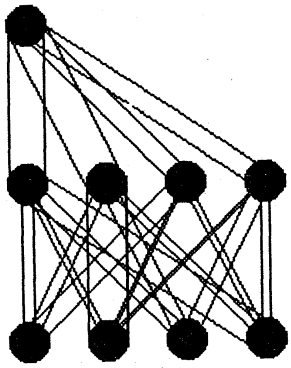

図 9 ＢＰ学習

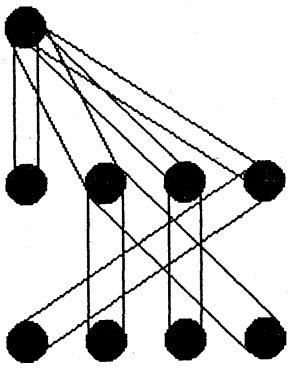

図 10 情報量を導入した学習

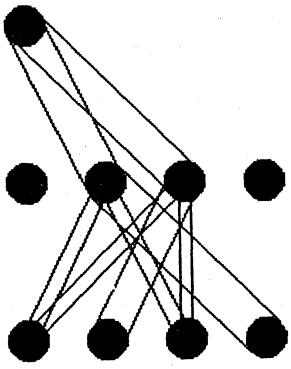

図 11 エントロピー最小化学習

5.おわりに

本研究では、ニューラルネットワークの構造の指標として情報量を提案した。情報量にはD_1、D_2、ＴＩがある。D_1はユニット単位の構造に関連し、D_2はユニット間の結合に関連する。ＴＩは両者を足したものでこれがネットワークの構造化の指標となる。また、その情報量を増やすことをＢＰ学習に導入し、ネットワークを構造化する学習法も提案した。計算機実験により、D_1を増やすことによりネットワークのユニット数を決定でき、ＴＩを増やすことによりネットワーク結合の構造化ができることを確かめた。

今後の課題として、情報量を導入する割合の決定の仕方、情報量を導入することによる収束性の解析、相互結合型ニューラルネットワークへの応用などがあげられる。

6.参考文献

［1］ Rumelhart D.E,McClelland J.L.："Parallel Distributed Processing"，MIT Press，Cambridge，Massachusetts，pp318-362,1986

［2］ 内田、石川：”エントロピー基準に基づいたニューラルネットワークの構造学習”、信学技報、ＮＣ９１－１５３（１９９１）

［3］ 芳我、石川：”各種構造学習法の構造化および汎化能力の比較”、信学技報、ＮＣ９２－２０（１９９３）