

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にくい。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、

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

となる。ここで、 f はシグモイド関数を用いる。

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

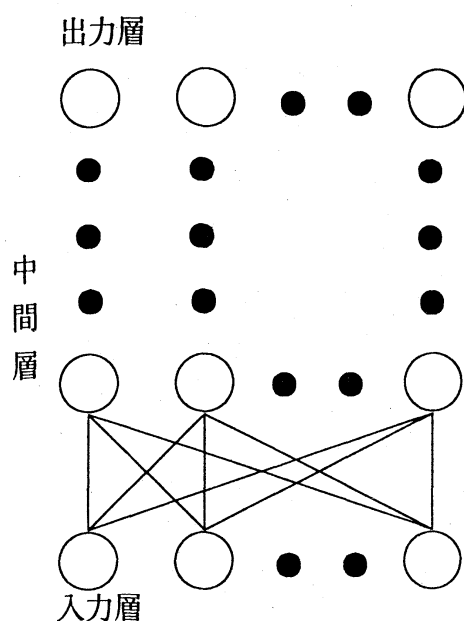


図 1 階層型ニューラルネットワーク

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

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

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、

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

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。

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w'_{ij} \quad (5)$$

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き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

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

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

とする。また、ユニット i が信号を受け取る確率 p_i を

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

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

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

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 $H_{1\max}$ 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。

$$D_1 = H_{1\max} - H_1 \quad (10)$$

$$H_1 = -\sum_{i=1}^N p_i \log p_i \quad (11)$$

$$H_{1\max} = \log N \quad (12)$$

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。

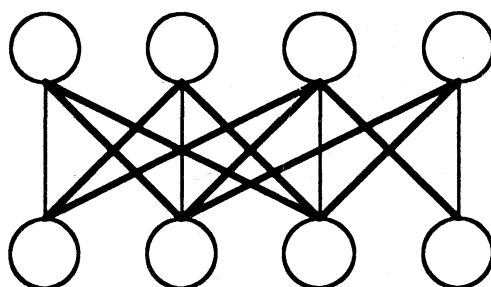


図 2 D_1 が小さいネットワーク

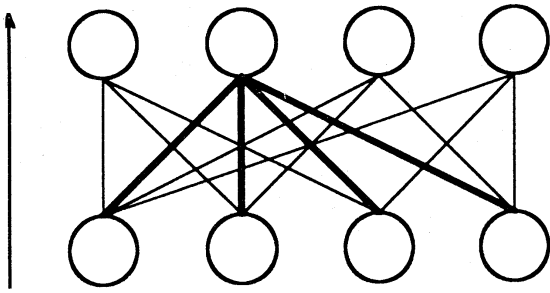


図3 D_1 が大きいネットワーク

(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。

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

$$H_M = -\sum_{i=1}^N \sum_{j=1}^N q_i p_{ij} \log p_{ij} \quad (14)$$

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。

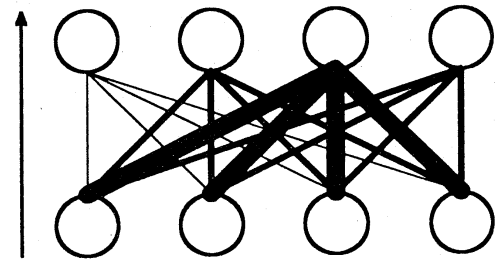


図4 D_2 が小さいネットワーク

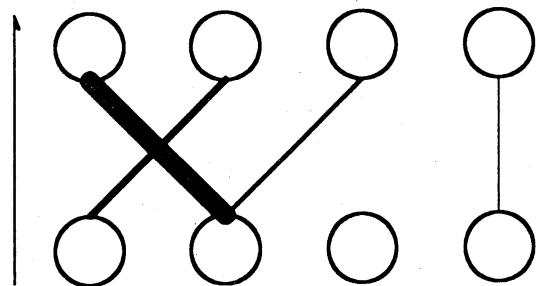


図5 D_2 が大きいネットワーク

(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。

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

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。

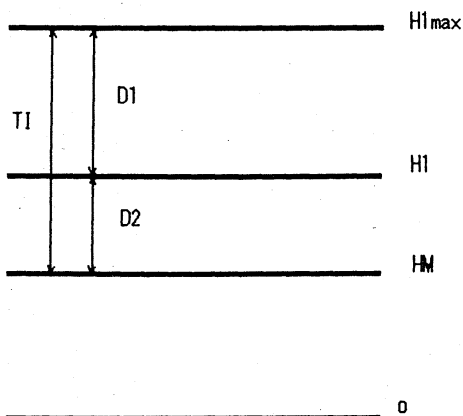


図 6 各情報量の関係

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 T_1 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

$$J = E - \lambda' D_1 \quad (16)$$

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

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

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ふ。そうすると、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ふ場合は $\lambda = 0$ 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 T_1 の導入

D_1 と同様に評価関数 J を

$$J = E - \lambda' T_1 \quad (18)$$

とする。 Δw_{ij} は

$$\begin{aligned} \Delta w_{ij} &= -\eta \frac{\partial J}{\partial w_{ij}} \\ &= -\eta \frac{\partial E}{\partial w_{ij}} + \eta \lambda' \frac{\partial T_1}{\partial w_{ij}} \\ &= -\eta \frac{\partial E}{\partial w_{ij}} - \lambda \frac{\partial H_M}{\partial w_{ij}} \quad (\lambda = \eta \lambda') \end{aligned} \quad (19)$$

となる。 D_1 と同様に微分値の符号と重みの符号が違ふ場合は $\lambda = 0$ 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

エントロピ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。

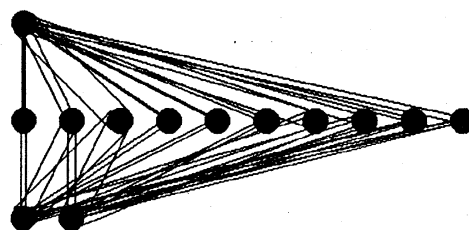


図7 BP学習によるネットワーク

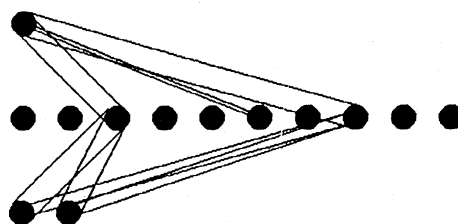


図8 情報量によるネットワーク

4.2. T I の最大化

入力層と中間層の間、中間層と出力層の間の T I 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。

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

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 B P 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。B P 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。

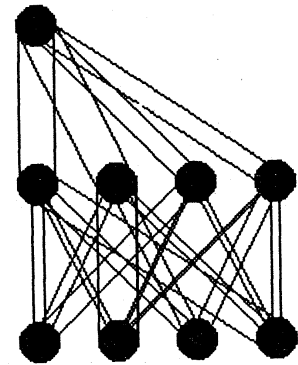


図 9 B P 学習

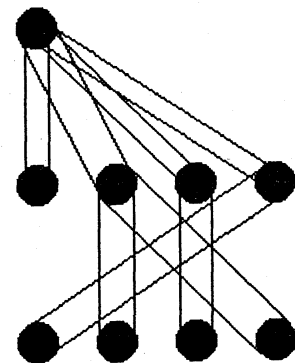


図 10 情報量を導入した学習

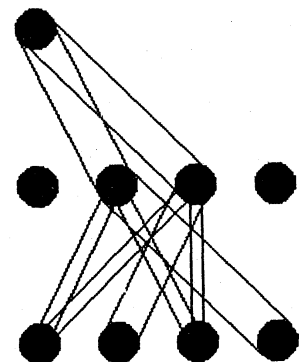


図 11 エントロピー最小化学習

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、NC 91-153 (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、NC 92-20 (1993)