研究論文

学習時間に制約がある環境に適した ニューラルネットワーク構造学習則*1

出村 公成*2, 中村 淳*3, 安西祐一郎*2

慶應義塾大学理工学部*2, 日立製作所*3

The Adaptive Network Architecture for a Time-constrained Environment

Kosei Demura*2, Atsushi Nakamura*3 and Yuichiro Anzai*2

Faculty of Science and Technology, Keio University*2, Hitachi, Ltd.*3

This paper proposes an adaptive network architecture called Hybrid SOLAR (Supervised One-shot Learning Algorithm for Real number inputs), which is a hybrid algorithm between a one-shot network construction algorithm and an iterative pruning algorithm. Hybrid SOLAR determines a network structure in two stages. In the first stage, Hybrid SOLAR requires only a single presentation of training examples to construct the network and learning is finished. In the second stage, the network prunes redundant weights to improve the generalization ability. Thus, Hybrid SOLAR retains the advantages of those two algorithms. It needs only a single presentation of the training set for learning and the generalization ability is satisfactory.

1. はじめに

誤差逆伝搬学習（以下、BP）を代表とする最急降下法に基づく学習則は、般化能力が高く工学的応用も容易なことから現在のところ最も広く使われているアルゴリズムであるが、Minsky や Fahlman ら¹⁾も指摘しているように、極小値問題、学習所要時間が大、ネットワーク構造やパラメータの決定が試行錯誤的であるなどといった問題があり、学習データが大規模化するほど問題は深刻になる。そのため、上記の問題を解決することはニューラルネットワーク研究における最も重要な課題の一つである。

上記の問題の中で主に試行錯誤的学習の解決及び般化能力を向上させるために、1980年代末からネットワーク構造学習則の研究が進められている。一般に、ネットワークの規模がデータに対して過小であると十分に学習できず、逆に過大であると般化能力が低下してしまうという問題がある。BPではネットワークの規模を試行錯誤的に決定していたが、構造学習則では高い般化能力を有するネットワークを構築するため、学

習例に適した規模のネットワークを学習則により決定する。構造学習則は大きく分けて、学習例を一度提示するだけでネットワーク構造を決定するワンショット構造学習則（以下、ワンショット学習則）と逐次的に構造を決定する逐次構造学習則がある。

ワンショット学習則は Namatame らにより Flash Learning 学習則として提案され²⁾、その後、出村らが入力に実数値を扱えるように Flash Learning を拡張した SOLAR (Supervised One-shot Learning Algorithm for Real number inputs)³⁾や般化能力を向上させた SOLAR2⁴⁾を提案している。ワンショット学習則は、学習例間の類似度の幾何学的構造に基づき識別関数となる超平面を決定する学習則であり、学習例を一度提示しただけでネットワーク構造及びパラメータを決定することができ、学習が終了する。そのため、最急降下法に基づいた学習則の大問題である極小値問題や試行錯誤的学習といった問題がなく、大規模なデータでも高速に学習が可能であり、学習に要する時間の予測も可能である。しかし、中間ユニットが局所的な情報表現になっているため逐次学習と比較すると必ずしも般化能力が高いとはいえず、ネットワークの規模が大きくなる傾向がある。

一方、逐次構造学習則には、小さなネットワーク構

*1 1995年10月3日受付

*2 〒223 横浜市港北区日吉 3-14-1

*3 〒319-12 茨城県日立市大みか町 7-1-1

造から学習を始め、必要なユニットを逐次追加して適切な規模のネットワーク構造を決定する生成的構造学習則と大きめのネットワーク構造から学習を始め、冗長な結合ユニットを逐次的に削除してネットワーク構造を決定する削除的構造学習則とがある。理論的にも、逐次構造学習則は実現する関数の複雑さと教師データに対する関数の精度とのトレードオフを決定する Structural Risk Minimization (SRM) 原理に基づくものであり、般化能力を高める方法であることを Vapnik が示している⁵⁾。

生成的構造学習則の代表的なものとして Mézard と Nadal の Tiling algorithm⁶⁾、Fahlman の Cascade-Correlation algorithm¹⁾ などの研究がある。一方、削除的構造学習則には、評価関数にペナルティ項を付け加えることにより学習をしながらネットワーク構造を決定する Chauvin⁷⁾ や石川⁸⁾ の研究と、最初に最急降下法などにより学習し、その後、結合の重要度に基づいて枝刈りを行う Sietsman ら⁹⁾、Le Cun らの Optimal Brain Damage¹⁰⁾ (以下 OBD) や Hassibi らの Optimal Brain Surgeon¹¹⁾ などの研究がある。

生成的学習則や評価関数にペナルティ項を付加する削除的構造学習則ではネットワーク構造が決定するまで学習は完了せず、ネットワークの構造が変化している間は正確な想起が困難であり般化能力も低く、極小値問題や学習速度が遅いので学習時間に制約のある環境下での使用は困難である。一方、結合の重要度に基づく削除的構造学習則の多くは、学習には最急降下法を使用するものが多いので最急降下法の問題点がある。

以上述べたように、ワンショット学習則と逐次構造学習則にはそれぞれ問題があり、各分野でこれらの問題を解決するために研究が進められている。しかし、各学習則の欠点はアルゴリズム固有の本質的な問題であるため解決するのは非常に困難であると考えられる。そこで、本研究では従来研究とは発想を変えて、ワンショット構造学習則と逐次構造学習則をハイブリット化することにより、上記の問題点を解決する構造学習則 Hybrid SOLAR を提案する。

Hybrid SOLAR では、まずワンショット学習則 SOLAR 2 により大規模なデータでも高速に学習を終了し、ネットワークを自動的に構築する。その後、般化能力を高めるために、削除的構造学習則を用いて、冗長な結合を逐次的に削除する。従って、学習時間に制約がある環境下でも、SOLAR 2 により高速に学習し、時間制約が解消された時点で般化能力の高いネットワークに再構築するので、実世界などの学習時間に

制約がある環境下での使用に適した学習則であると考えられる。

2. SOLAR2

2.1 定 義

ネットワーク構造として入力ユニット数 N 個、中間ユニット数 K 個、各出力ユニットは独立なので出力ユニット数 1 個の 3 層ネットワークを考える。

ネットワークへの p 番目の入力データ $\mathbf{x}_p$ は次に示す N 個の属性値をもつ N 次元行ベクトルである。

$$\mathbf{x}_p = [x_{p1}, x_{p2}, \dots, x_{pN}], \quad 0 \leq x_{pi} \leq 1 \quad (1)$$

ここで、 x_{pi} は $\mathbf{x}_p$ の i 番目の要素である。

入力データ $\mathbf{x}_p$ により、全入力データ $\mathbf{X}$ は以下のように表示される。

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p, \dots, \mathbf{x}_P]^t \quad (2)$$

但し、 $\mathbf{x}_p \neq \mathbf{x}_i$ ($1 \leq p \neq i \leq P$)

ここで、 P は全入力データ数、 t は転置行列を表す。

中間ユニット j の全出力値 (以下、中間概念) $\mathbf{H}_j$ は中間ユニット j の出力 h_{pj} により以下で示される P 次元列ベクトルである。

$$\mathbf{H}_j = [h_{1j}, h_{2j}, \dots, h_{Pj}]^t, \quad h_{pj} \in \{0, 1\} \quad (3)$$

入力データ $\mathbf{x}_p$ に対応する教師データの値を c_p とする。全教師データ (以下、概念) $\mathbf{C}$ は教師データ c_p により次式で示される P 次元列ベクトルである。

$$\mathbf{C} = [c_1, c_2, \dots, c_p, \dots, c_P]^t, \quad c_p \in \{0, 1\} \quad (4)$$

学習データ $(\mathbf{x}_p, c_p)$ は正事例 C^+ または負事例 C^- であり次式で定義される。

$$C^+ = \{\mathbf{x}_p \in \mathbf{X} : c_p = 1\} \quad (5)$$

$$C^- = \{\mathbf{x}_p \in \mathbf{X} : c_p = 0\} \quad (6)$$

2.2 アルゴリズム

定義 1 全入力データ $\mathbf{X}$, 概念 $\mathbf{C}$ における入力データ $\mathbf{x}_p$ と $\mathbf{x}_q$ の動的類似度 $s_a(\mathbf{x}_p, \mathbf{x}_q, \mathbf{C})$ を次式で定義する。

Fig. 1 Network structure.

$$s_a(\mathbf{x}_p, \mathbf{x}_q, \mathbf{C}) = \sum_{i=1}^N a_i x_{pi} x_{qi}, \quad (7)$$

$$a_i = \frac{r_i}{r_{\max}}, \quad r_i = \left\| \frac{\sum_{p=1}^P (x_{pi} - \bar{I}_i)(c_p - \bar{C})}{\sqrt{\sum_{p=1}^P (x_{pi} - \bar{I}_i)^2} \sqrt{\sum_{p=1}^P (c_p - \bar{C})^2}} \right\| \quad (8)$$

ここで $\sqrt{\sum_{i=1}^N a_i x_{pi}^2} = 1$ ($p, q=1, 2, \dots, P$) と正規化する。 $\mathbf{I}_i = [x_{1i}, x_{2i}, \dots, x_{pi}, \dots, x_{Pi}]^t$ であり、 $\bar{C}$ 及び $\bar{I}_i$ はそれぞれ、 $\mathbf{C}$ 及び $\mathbf{I}_i$ の平均値、 $r_{\max}$ は r_i の最大値である。

定義2 類似度行列 $\mathbf{D}(\mathbf{X}, \mathbf{C})$ は、次式で示す $P \times 2$ 行列である。

$$\mathbf{D}(\mathbf{X}, \mathbf{C}) = \mathbf{S}(\mathbf{X}, \mathbf{C})[\mathbf{C}, \mathbf{1} - \mathbf{C}] \quad (9)$$

$$= \begin{bmatrix} G^+(\mathbf{x}_1), & G^-(\mathbf{x}_1) \\ G^+(\mathbf{x}_2), & G^-(\mathbf{x}_2) \\ \vdots & \vdots \\ G^+(\mathbf{x}_P), & G^-(\mathbf{x}_P) \end{bmatrix} \quad (10)$$

対称行列 $\mathbf{S}(\mathbf{X}, \mathbf{C})$ は $P \times P$ 行列であり、 pq 成分は式(7)で示される。 $\mathbf{D}(\mathbf{X}, \mathbf{C})$ の p 行 1 列成分である $G^+(\mathbf{x}_p)$ は、次式で示される入力データ $\mathbf{x}_p$ と正事例全体との平均類似度であり、 $\mathbf{D}(\mathbf{X}, \mathbf{C})$ の p 行 2 列成分である $G^-(\mathbf{x}_p)$ は $\mathbf{x}_p$ と負事例全体との平均類似度である。

$$G^+(\mathbf{x}_p) = \frac{\sum_{q=1}^P s_a(\mathbf{x}_p, \mathbf{x}_q, \mathbf{C}) c_q}{\sum_{q=1}^P c_q} \quad (11)$$

$$G^-(\mathbf{x}_p) = \frac{\sum_{q=1}^P s_a(\mathbf{x}_p, \mathbf{x}_q, \mathbf{C})(1 - c_q)}{\sum_{q=1}^P (1 - c_q)} \quad (12)$$

定理1 類似度行列が線形分離可能である場合は、次式を満たす α, β, θ が存在する。

$$\begin{cases} \mathbf{x}_p \in \mathbf{C}^+ \text{ ならば } \alpha G^+(\mathbf{x}_p) + \beta G^-(\mathbf{x}_p) + \theta > 0 \\ \mathbf{x}_p \in \mathbf{C}^- \text{ ならば } \alpha G^+(\mathbf{x}_p) + \beta G^-(\mathbf{x}_p) + \theta < 0 \end{cases} \quad (13)$$

すなわち、類似度平面は

$$\alpha G^+(\mathbf{x}_p) + \beta G^-(\mathbf{x}_p) + \theta = 0 \quad (14)$$

により、正事例及び負事例の領域に分割される。線形分離可能な場合は、中間層のない2層ネットワークとなり、式(14)に式(11)、(12)を代入すると、出力ユニットの活性値 U を次式で求めることができる。

$$U(\mathbf{x}_p) = \sum_{i=1}^N w_i x_{pi} + \theta \quad (15)$$

ここで、結合荷重 w_i は次式で与えられる。

Fig. 2 Step function.

$$w_i = \alpha \frac{\sum_{q=1}^P a_i x_{qi} c_q}{\sum_{q=1}^P c_q} + \beta \frac{\sum_{q=1}^P a_i x_{qi} (1 - c_q)}{\sum_{q=1}^P (1 - c_q)} \quad (16)$$

各ユニットの出力関数 f は次式と Fig. 2 に示すステップ関数である。

$$\begin{cases} U \geq 0 \text{ ならば } f(U(\mathbf{x}_p)) = 1 \\ U < 0 \text{ ならば } f(U(\mathbf{x}_p)) = 0 \end{cases} \quad (17)$$

また、出力ユニットの出力 O は次の性質をもつ。

$$\begin{cases} \mathbf{x}_p \in \mathbf{C}^+ \text{ ならば } O = f(U(\mathbf{x}_p)) = 1 \\ \mathbf{x}_p \in \mathbf{C}^- \text{ ならば } O = f(U(\mathbf{x}_p)) = 0 \end{cases} \quad (18)$$

以上は線形分離可能な場合のアルゴリズムである。線形分離不可能な場合は、概念 $\mathbf{C}$ を線形分離可能な中間概念 $\mathbf{H}_j$ に次式のように分割することにより線形分離可能問題に帰着させる。

$$\mathbf{C} = \bigoplus_{j=1}^K \mathbf{H}_j \quad (19)$$

但し、 $\mathbf{H}_i \cdot \mathbf{H}_j = \mathbf{0}$, $\mathbf{H}_i \neq \mathbf{0}$ ($i \neq j$: $i, j=1, 2, \dots, K$) ここで $\bigoplus$ は列ベクトル $\mathbf{H}_j$ ($1 \leq j \leq K$) のビットごとの論理和、 $\mathbf{0}$ は学習データ数が P 個の場合、 P 次元の0列ベクトルを表す。分割の具体的な方法については文献³⁾を参考にされたい。

3. Hybrid SOLAR

3.1 ハイブリッド化の基本的な考え方

ワンショット構造学習則と逐次構造学習則をハイブリッド化するための基本的な考え方について述べる。Hybrid SOLAR では、SOLAR 2 を用いて学習例の一回提示によりネットワークを構築する。SOLAR 2 により構築されるネットワーク構造は Fig. 1 に示すような3層階層型ネットワークである。入力ユニット i から中間ユニット j への結合荷重 w_{ji} 及び中間ユニットから出力ユニットへの結合荷重 v_j は次式である。

$$w_{ji} = \alpha \frac{\sum_{q=1}^P a_i x_{qi} h_{qj}}{\sum_{q=1}^P h_{qj}} + \beta \frac{\sum_{q=1}^P a_i x_{qi} (1 - h_{qj})}{\sum_{q=1}^P (1 - h_{qj})}$$

$$v_j = 1$$

式(19)の概念 C 及び中間概念 H_j をネットワーク構造で考えると、それぞれ出力ユニット値及び中間ユニット値に対応する。したがって、出力ユニット値は各中間ユニット値の論理和であるので、中間ユニットから出力ユニットの結合荷重 v_j は 1 に固定されている。SOLAR 2 では類似行列が線形分離不可能な場合、中間概念が線形分離可能になるまで概念を分割して中間概念すなわち中間ユニットを生成する。従って、非線形性が強いほど中間ユニット数は多数生成され、中間ユニットは局所的な表現となる。そのため、最急降下法に基づく学習則と比較するとネットワークの規模は大きくなり般化能力は低くなる。

そこで、本研究では SOLAR 2 によって構築された局所的な表現のネットワークを削除的構造学習則により分散的な表現を持つ般化能力の高いネットワークに再構築する。Hybrid SOLAR では削除的構造学習則として OBD を用いるが、SOLAR 2 で構築されたネットワークに直接 OBD を適用することはできない。何故なら、OBD は結合を削除する基準として結合荷重に関する自乗誤差和の 2 階偏微分を結合荷重の重要度 (saliency) と定義し、重要度の一番小さな結合から削除する。

そのため、SOLAR 2 により構築されたネットワークに OBD を適用するために以下のようにネットワークを変更する。

第一に SOLAR 2 では中間ユニット j から出力ユニットへの結合荷重 v_j は 1 に固定されていたが、結合荷重 v_j を可変にする。第二に、各ユニットの出力関数 $f(x)$ を Fig. 2 に示すステップ関数から Fig. 3 及び次式に示すシグモイド関数に変更する。

Fig. 3 Sigmoid function.

$$f(x) = \frac{1}{1 + e^{-A(t)x}}$$

$$A(t) = \frac{\nu}{t} + 1$$

(20)

ここで、 t は訓練回数、 ν は実定数である。 $A(t)$ を導入することにより、学習開始時は $t=0$ であるから $A(t)$ は無限大になるので出力関数 $f(x)$ はステップ関数になり、訓練回数が増加するにつれて出力関数 $f(x)$ の傾きがなだらかになり $f(x)$ は $1/(1+e^{-x})$ に近づく。

この変更により、SOLAR 2 によりワンショットで構築された局所的な表現であるネットワークに OBD の手法を適用することが可能となり、分散表現である般化能力の高いネットワークに再構築することができる。

3.2 削除的構造学習則の適用

SOLAR 2 により構築されたネットワークに削除的構造学習則を適用する方法について具体的に述べる。本論文では、削除的構造学習則として OBD を一部改良したものを用いる。まず、OBD について述べる。

OBD は結合を削除する基準として結合荷重に関する自乗誤差和の 2 階偏微分を結合荷重の重要度 (saliency) と定義し、重要度の一番小さな結合から削除する。

SOLAR 2 で生成されたネットワークは Fig. 1 で示されるように、各中間ユニットは出力ユニットとは 1 個しか結合していないので、OBD に基づいて計算すると以下ようになる。まず、入力データ x_p に対する自乗誤差和 E_p を w_{ji} に関して Taylor 展開すると次式が得られる。

$$E_p(\mathbf{w}) = \sum_j \sum_i \frac{\partial E_p}{\partial w_{ji}} w_{ji} + \frac{1}{2} \sum_j \left(\sum_i \frac{\partial^2 E_p}{\partial w_{ji}^2} w_{ji}^2 \right) + \frac{1}{2} \sum_j \left(\sum_{i \neq i'} \frac{\partial^2 E_p}{\partial w_{ji} \partial w_{ji'}} w_{ji} w_{ji'} \right) + O(\|\mathbf{w}\|^3) \quad (21)$$

式(21)の第 1 項の $\frac{\partial E_p}{\partial w_{ji}}$ は E_p が極小値ならば 0 であり、第 4 項は高次のため無視する。また、計算量を減少させるため、結合荷重同士の干渉がないと仮定する (以下、対角化) と第 3 項は 0 になり、第 2 項だけが残る。したがって、結合荷重 v_j , w_{ji} の評価関数 $E(v_j)$, $E(w_{ji})$ は次式になる。

$$E(v_j) = \sum_p \frac{1}{2} \frac{\partial^2 E_p}{\partial v_j^2} v_j^2 \quad (22)$$

$$E(w_{ji}) = \sum_p \frac{1}{2} \frac{\partial^2 E_p}{\partial w_{ji}^2} w_{ji}^2 \quad (23)$$

また、結合荷重 v_j , w_{ji} に関する E_p の 2 階偏微分は、

入力ユニット i への入力 x_{pi} , 中間ユニット j の出力 h_{pj} , 出力ユニット o の出力 o_p , 教師データ c_p を用いて次のように表せる.

$$\begin{aligned}\frac{\partial^2 E_p}{\partial v_j^2} &= \frac{\partial^2 E_p}{\partial o_p^2} \left(\frac{\partial o_p}{\partial v_j} \right)^2 + \frac{\partial E_p}{\partial o_p} \frac{\partial^2 o_p}{\partial v_j^2} \\ &= A^2(t) h_{pj}^2 \sigma_p \{ \sigma_p - (c_p - o_p)(1 - 2o_p) \} \quad (24) \\ \frac{\partial^2 E_p}{\partial w_{ji}^2} &= \frac{\partial^2 E_p}{\partial h_{pj}^2} \left(\frac{\partial h_{pj}}{\partial w_{ji}} \right)^2 + \frac{\partial E_p}{\partial h_{pj}} \frac{\partial^2 h_{pj}}{\partial w_{ji}^2} \\ &= A^4(t) \sigma_p \{ \sigma_p - (c_p - o_p)(1 - 2o_p) \} v_j^2 \cdot \delta_{pj}^2 x_{pi}^2 \\ &\quad - A^3(t) (c_p - o_p) \sigma_p v_j \cdot \delta_{pj} (1 - 2h_{pj}) x_{pi}^2 \quad (25)\end{aligned}$$

ただし, $\sigma_p = o_p(1 - o_p)$, $\delta_{pj} = h_{pj}(1 - h_{pj})$
枝切りは式(22), (23)を計算して重要度の一番小さな結合を一つだけ削除する. その後, BP により再学習を行ない結合荷重を修正する. 評価関数が決められた目標値に収束しなくなるまで, OBD による枝切りと BP による再学習を繰り返すことにより中間ユニットが局所的な表現から徐々に分散的な表現に般化能力の高いネットワークを構築することが可能となる.

以上が OBD のアルゴリズムであるが, OBD では saliency の一番小さな結合を削除するわけであるが, saliency が一番小さな結合の結合荷重が必ずしも 0 に近いとは限らない. また, 計算量を減少させるため対角化の仮定を用いているが, その仮定は必ずしも正しくないため, OBD により選ばれた結合がネットワークにとって最も重要度が小さいとは限らない. 従って, 間違った枝を切ることにより般化能力が減少する場合がある. そのため, ここでは saliency の一番小さな結合をすぐ削除するのではなく, 以下のように徐々に結合荷重を減少させ, 十分に 0 に近付いたら結合を削除するようにアルゴリズムを変更する.

$$w_{new} = \gamma w_{old} \quad (26)$$

ここで, w_{old} , w_{new} はそれぞれ更新前及び更新後の結合荷重, $0 < \gamma < 1.0$ とする. 以下, 本論文では OBD を一部改良したものを OBD と記す.

3.3 アルゴリズム

Hybrid SOLAR のアルゴリズムをまとめると以下のようなになる.

- ステップ 1 SOLAR 2 によりワンショットでネットワークを構築する.
- ステップ 2 結合荷重の重要度を計算する.
- ステップ 3 結合荷重を保存する.
- ステップ 4 ステップ 2 で計算した結合荷重の重要度のうちで最小値の結合荷重を $w_{new} = \gamma w_{old}$ と更新し, 十分に 0 に近付いたら結合を削除する. ここで, γ は 0 から 1 までの定数である.

る.

ステップ 5 誤差逆伝搬学習により再学習を行なう. 評価関数が決められた目標値以下となる場合はステップ 2 へ戻る. その他の場合はステップ 2 で保存していた結合荷重の値を用いアルゴリズムを終了する.

4. シミュレーション

Fisher に紹介されて以来, 多変量解析における判別問題の典型的な問題であるアヤメの分類問題に本手法を適用する. Fisher のアヤメのデータは, ヴァージニカ, ヴェルシカラー及びセトーサの 3 種類のアヤメ (各種類 50 サンプル, 合計 150 サンプル) から 4 個の属性, がく片の長さ, がく片の幅, 花びらの長さ及び花びらの幅を計測したデータである.

まず, 般化能力のテストとして, 学習データ数に対する分類の正解率を Hybrid SOLAR, SOLAR 2, BP, BP に OBD を用いた学習則 (BP+OBD) の 4 つのアルゴリズムに対して調べた.

学習には学習データ数を 0 個から 150 個まで 5 個ずつ変化させ, 各場合について正解率を調べた. Fig. 4 はその結果を示してあり, 横軸は学習データ数, 縦軸は未知のデータに対する正解率である. 例えば, データ数が 50 個の場合は, 全データ 150 個の中からランダムにデータを 50 個選び学習させ, 残りの学習していない 100 個のデータに対する正解率を調べている. ネットワークへの入力アヤメの属性値なので入力ユニット数は 4 個, 出力はアヤメの種類に対応しているので出

Fig. 4 Success rate for unknown data.

カユニット数は3個である。BPの中間ユニット数を4個、BP+OBDでは中間ユニット数を7個から削除的学習を開始させた。BPの学習の際は、学習係数を1.0、慣性項を0.2に設定し、評価関数(平均自乗誤差和)が0.01になるまで学習を行なった。一方、SOLAR 2は学習により中間ユニットが17個生成し、Hybrid SOLARではSOLAR 2が生成したネットワーク構造を初期ネットワークとし、式(20)の減衰率 γ を1000とした。Fig. 4の結果を見ると明らかに、SOLAR 2ネットワークは局所的な表現のため正解率は最高約80%と他の学習則と比較して般化能力は低い。一方、Hybrid SOLARではOBDを用いて分散的な表現へとネットワーク構造を変換しているためBPやBP+OBDと同程度の正解率を示し高い般化能力を有する。BP+OBDと比較してHybrid SOLARの正解率が若干低い理由としては、Hybrid SOLARでは、各中間ユニットは出力ユニット1個としか結合していないため、各中間ユニットが全ての出力ユニットと結合しているBPネットワークと比較するとネットワーク構造が局所的になっているためだと考えられる。したがって、Hybrid SOLARにおいても中間ユニットが複数の出力ユニットと結合するように拡張すると般化能力がさらに向上するものと考えられる。

次に、訓練回数と結合の削除率の関係について調べ

る。この結果はHybrid SOLARのネットワーク構造がどの程度局所的な構造から分散的な構造に変化しているかについて示唆を与えるものであり、SOLAR 2で構築されたネットワークにOBDを適用する有用性がわかる。Fig. 5は学習データ数25個、Fig. 6は学習データ数50個と75個の場合の結果であり、全データ150個からランダムにデータセットを30回選び平均を示している。横軸は訓練回数、縦軸はネットワークの全結合数である。データ数25の場合は式(26)の $\gamma_1=0.9$ (入力層から中間層間)、 $\gamma_2=0.2$ (中間層から出力層間)とし、データ数50の場合は $\gamma_1=0.6$ 、 $\gamma_2=0.2$ 、データ数75の場合は $\gamma_1=0.5$ 、 $\gamma_2=0.1$ とした。最適な γ を求める方法はないが、一般的にはデータ数が少ない時には γ を大きくしないとOBDの対角化による誤差のため枝刈りが最後まで行なわれず、ネットワーク構造が十分小さくならない場合がある。

Fig. 5では、訓練回数0のときは、SOLAR 2で生成された結合数は平均87であり、枝刈り終了後は結合数は平均32となり、約37%まで結合数が減少した。一方、BPでは当初結合数49から枝刈りを始め、枝刈り終了後は、結合数は平均19まで減少した。次に、Fig. 6を見ると明らかに、Hybrid SOLARではデータ数が多くなるにつれ、初期のネットワーク構造が大きくなることわかる。しかし、OBDを適用すること

Fig. 5 Number of connections versus training epochs for Hybrid SOLAR and BP+OBD. The number of training data is 25.

Fig. 6 Number of connections versus training epochs for Hybrid SOLAR.

により学習終了後は、ほぼ同程度の規模のネットワーク構造まで収束された。以上、BPと比較するとSOLAR 2で生成されたネットワークの規模は大きいのでOBDによる枝刈りはネットワークの規模を小さくするうえで重要な働きをしており、Fig. 4の結果とあいまって削除的学習が非常に有効に作用していることがわかる。

更に、枝刈りが終了するまでの訓練回数についても比較した。Table 1は学習データ数 25, 50, 75 個についてHybrid SOLARとBP+OBDとの訓練回数を比較した結果を示している。表を見ると明らかなように学習データ数 25 個の場合はHybrid SOLARの方が約 2.3 倍、データ数 50 個の場合は約 4.0 倍、データ数 75 個の場合は約 4.9 倍、訓練回数が多い。これは、SOLAR 2では学習例を一度提示するだけで高速に学習できるが、SOLAR 2により生成された初期ネットワークが学習データ数の増加にともないBP+OBDより大きくなるためである。

従って、一般的にはBPと比較してHybrid SOLARの学習に要する訓練回数は少ないとはいえない。しかしながら、Hybrid SOLARではSOLAR 2によりOne-shotで学習が終了し、学習データに対しては100%の正解率であり、OBDに移行したところでアルゴリズム ステップ3にあるように常に結合荷重を保存しているので、学習の途中でも学習データに対しては常に正しい答を出力することができる。

最後に、削除的学習によりネットワークの分散表現がどのように変化するかを示す。Fig. 7は学習データ数 25 個の場合におけるネットワークの分散表現の程度を定量的に示した図である。横軸は訓練回数、縦軸は入力データ x_p に対する中間ユニット j の出力 h_{pj} の分散 $\delta^2 = \sum_{p=1}^{150} \sum_{j=1}^K (h_{pj} - \bar{h})^2 / 150K$ である。ここで、 $\bar{h}$ は全中間ユニットの平均出力、 K は各訓練回数に対応する中間ユニット数である。分散が小さいほど中間ユニットが少数のデータにのみ反応して出力するのでネットワークは局所表現的であることを示し、逆に、分散が大きいほど多数のデータに反応して出力するので分散的であることを示している。本論文では、中間ユニットの出力の分散をネットワークの分散表現の尺度とする。Fig. 7を見ると明らかなように訓練回数 1000 回くらいまでに分散が急激に上昇し、それ以降はゆるやかに分散が上昇している。すなわち、SOLAR 2により生成された局所表現的なネットワークは訓練回数 1000 回くらいまでのうちにOBDにより急激に分散表現的なネットワークに変換され、それ以降はゆるやかに変化していることがわかる。

Table 1 Training epochs (learning rate=1.0, momentum term=0.2).

Networks	No. of training data	No. of initial weights	Epochs
Hybrid SOLAR	25	87	9,472
	50	169	17,059
	75	243	19,260
BP+OBD	25	49	4,105
	50	49	4,230
	75	49	3,950

Fig. 7 Variance of hidden units versus training epochs when the number of training data is 25.

5. おわりに

本論文では、ワンショット学習則と逐次構造学習則をハイブリッド化することにより、学習時間に制約がある環境に適した般化能力の高いニューラルネットワーク学習則 Hybrid SOLAR を提案した。

ワンショット構造学習則は、学習例を一度提示するだけでネットワーク構造が決定し、学習も終了するので学習速度が非常に速いが、般化能力は必ずしも高くない。また、逐次構造学習則は、般化能力は高いが、極小値問題や学習時間が遅いといった問題がある。そこで、Hybrid SOLARでは両者をハイブリッド化することにより、それらの問題点を解決している。Hybrid SOLARはワンショット学習で高速に学習が可能であり、そのうえ、高い般化能力を有している。学習時間に制約がある環境下での使用に適した学習則であると

考えられる。

今後の課題としては、ワンショット学習則により生成したネットワークに削除的学習を適用した初期の段階で自乗誤差和が増加する問題、枝を切り過ぎることによる般化能力が低下する問題の解決及び中間ユニットが複数の出力ユニットと結合するように学習則を拡張することがあげられる。

謝 辞 本研究の一部は、文部省科学研究費重点領域研究「脳の高次処理」# (05267103) 及び日本学術振興会からの援助による。最後に有益なコメントを頂いた慶應義塾大学 安西研 全炯浩氏に感謝します。

参 考 文 献

- 1) Fahlman, S.E. and Lebiere, C. (1990): The cascade-correlation learning. In *Advances in Neural Information Processing Systems 2*. (pp. 524-532) San Mateo: Morgan Kaufmann Publishers
- 2) Namatame, A. and Tsukamoto, Y. (1992): Structural connectionist learning with complementary coding. *International Journal of Neural Systems*, Vol. 3, No. 1, pp. 19-31
- 3) 出村公成, 梶浦正浩, 安西祐一郎 (1994): 実数入力データを一回の提示により学習できる教師あり学習則. *電子情報通信学会論文誌*, Vol. J77-D-II, No. 10, pp. 2083-2092
- 4) 出村公成, 安西祐一郎 (1995): 動的類似度を用いたワンショット学習則. *日本神経回路学会誌*, Vol. 2, No. 1, pp. 4-10
- 5) Vapnik, V. N. (1995): *The Nature of Statistical Learning Theory*. (pp. 89-100) New York: Springer
- 6) Mézard, J. and Nadal, J. (1989): Learning in feedforward layered networks: the tiling algorithm. *Journal of Physics A*, Vol. 22, pp. 2191-2203
- 7) Chauvin, Y. (1988): A back-propagation algorithm with optimal use of hidden units. In *Advances in Neural Information Processing Systems 1*. (pp. 519-526) San Mateo: Morgan Kaufmann Publishers
- 8) 石川真澄 (1992): コネクショニストモデルの忘却を用いた一般構造学習. *電気学会論文誌 C*, Vol. 112, No. 6, pp. 365-369
- 9) Sietsman, J. and Dow, R. J. F. (1988): Neural net pruning—why and how. *IEEE International Conference on Neural Networks*, Vol. 1, pp. 325-333
- 10) Le Cun, Y., Denker, S. and Solla, A. (1990): Optimal brain damage. In *Advances in Neural Information Processing Systems 2*. (pp. 598-605) San Mateo: Morgan Kaufmann Publishers
- 11) Hassibi, B. and Stork, D. G. (1993): Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5*. (pp. 164-171) San Mateo: Morgan Kaufmann Publishers