計測自動制御学会東北支部 第215回研究集会 (2004.5.27)
資料番号 215-4

# 共役勾配法に基づく学習アルゴリズムの挙動解析

# Behavior Analysis of a Learning Algorithm based on the Conjugate Gradient Method

西山 清*，○渡辺 充範*

Kiyoshi NISHIYAMA*, ○ Mitsunori WATANABE*

*岩手大学工学部情報システム工学科

*Dep. of Computer & Infomation Science, Iwate University

**キーワード**： ニューラルネットワーク (neural network), 学習アルゴリズム (learning algorithm), 共役勾配法 (conjugate gradient method), 直線探索 (line-search), 誤差曲面 (error surface)

**連絡先**： 〒020-8551 盛岡市上田4-3-5 岩手大学 工学部 情報システム工学科
西山 清, Tel.: 019-621-6475，Fax.: 019-621-6475，E-mail: nisiyama@cis.iwate-u.ac.jp

## 1. はじめに

ニューラルネットワークにおける学習とは、所望の入出力関係 (写像) を満たすようにニューロン間の結合重みとしきい値を更新する過程といえる。

本研究では、排他的論理和 (XOR) 問題を用いて、共役勾配法に基づく学習アルゴリズムの挙動を重み空間と誤差曲面において詳細に解析する。

## 2. 共役勾配法

ニューラルネットワークの学習は教師データと非線形なシステムにおいて得られた出力との誤差を最小とする重みを求める問題であるため非線形最適化問題とみなすことができる。そこで、非線形最適化問題の解法の1つである共役勾配法をニューラルネットワークの学習に適用する。

図1に共役勾配法における重み更新の方向を示す。

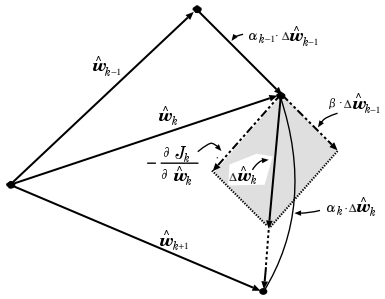



Fig. 1　共役勾配法の重み更新

共役勾配法を適用する前に、多層ニューラルネットワークを線形化した次の状態空間モデルで表現する。

$$\begin{aligned} \boldsymbol{w}_{k+1} &= \boldsymbol{w}_k \\ m_k &= \boldsymbol{H}_k\boldsymbol{w}_k + v_k,\quad z_k = \boldsymbol{H}_k\boldsymbol{w}_k \end{aligned}$$

ただし、

$$\begin{aligned} m_k &= y_k - h_k(\hat{\boldsymbol{w}}_{k|k-1}) + \boldsymbol{H}_k\hat{\boldsymbol{w}}_{k|k-1} \\ \boldsymbol{H}_k &= \left.\frac{\partial h_k(\boldsymbol{w})}{\partial \boldsymbol{w}}\right|_{w=\hat{w}_{k|k-1}} \end{aligned}$$

次に、目的関数を $J = (m_k - \boldsymbol{H}_k\hat{\boldsymbol{w}}_k)^2$ として共

役勾配法を適用すれば次の学習アルゴリズムが得られる。

**[共役勾配法に基づく学習アルゴリズム]**

$$
\begin{aligned}
\hat{\boldsymbol{w}}_{k+1} &= \hat{\boldsymbol{w}}_k + \alpha_k \cdot \Delta\hat{\boldsymbol{w}}_k \\
\Delta\hat{\boldsymbol{w}}_k &= -\frac{\partial J_k}{\partial \hat{\boldsymbol{w}}_k} + \beta \cdot \Delta\hat{\boldsymbol{w}}_{k-1}, k = 0, 1, 2, \cdots
\end{aligned}
$$

ただし、

$$
\begin{aligned}
\Delta\hat{\boldsymbol{w}}_0 &= -\frac{\partial J_0}{\partial \hat{\boldsymbol{w}}_0} \\
\alpha_k &: \text{直線探索によって求める} \\
\beta &: \text{経験的に決定する} \\
J_k &= (m_k - \boldsymbol{H}_k\hat{\boldsymbol{w}}_k)^2 \\
&= \hat{\boldsymbol{w}}_k^T \boldsymbol{H}_k^T \boldsymbol{H}_k \hat{\boldsymbol{w}}_k - 2m_k \boldsymbol{H}_k \hat{\boldsymbol{w}}_k + m_k^2 \\
\frac{\partial J_k}{\partial \hat{\boldsymbol{w}}_k} &= -2\boldsymbol{H}_k(m_k - \boldsymbol{H}_k\hat{\boldsymbol{w}}_k) \\
m_k &= y_k - h_k(\hat{\boldsymbol{w}}_k) + \boldsymbol{H}_k\hat{\boldsymbol{w}}_k
\end{aligned}
$$

**直線探索アルゴリズム (等間隔 V 型 3 点探索法)** $J(w)$ は唯一の最小点をもつ連続な 1 変数関数とする。図 2のように、$\hat{w}_0 < \hat{w}_1 < \hat{w}_2$、$J(\hat{w}_0) > J(\hat{w}_1) < J(\hat{w}_2)$ となっている 3 点 $(\hat{w}_0, \hat{w}_1, \hat{w}_2)$ を V 型3 点という。このような V 型3 点を見い出せば、$J(w)$ の最小点は必ず、区間 $[\hat{w}_0, \hat{w}_2]$ の間にあることは確かである。

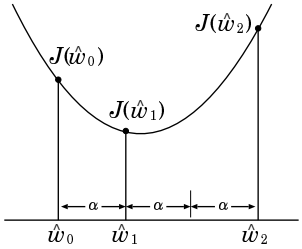



Fig. 2　V 型 3 点

従って、$J(w)$ の最小点を定めるには、まずこの V 型 3 点を見い出すことが必要となる。初期点 $\hat{w}_0$を適当にとり、探索のステップ長さを$\alpha_0$として、ある程度粗い探索を行って V 型 3 点を見い出す。まず $J(\hat{w}_0)$ を求め、$\hat{w}_1 = \hat{w}_0 + \alpha$とし $J(\hat{w}_1)$ を求める。$J(\hat{w}_0) > J(\hat{w}_1)$ ならば踏み出した方向は正しい。もし $J(\hat{w}_1) > J(\hat{w}_0)$ ならば、踏み出した方向は誤りであるから、$\alpha$の符号を反対にしてやり直す。次に、ステップを倍にして$\hat{w}_2 = \hat{w}_1 + 2\alpha$とし $J(\hat{w}_2)$ を求める。ここで $J(\hat{w}_2) < J(\hat{w}_1)$ ならば、さらにステップを倍にして$\hat{w}_3$を決め、同じことを繰り返す。$J(\hat{w}_2) > J(\hat{w}_1)$ ならば、V 型 3 点が見い出されたことになる。この方法で V 型 3 点が見い出されると一般にそれは図 2に示すように、$\hat{w}_{i+1}$は $[\hat{w}_i, \hat{w}_{i+2}]$ を 1:2 に内分する点となっている。ここで、$\hat{w}_{i+1}$と$\hat{w}_{i+2}$の中点で $J(w)$ の値を求め $J(\hat{w}_{i+1})$ と比較し、値の小さいほうを等間隔 V 型 3 点の最小点とする。次にその最小点を初期値として、たとえば $\alpha = \alpha/10$ と置き、上記のアルゴリズムを繰り返す。間隔が十分小さくなったとき $(\alpha < \alpha_{solv})$ に停止する。また、最小点がない場合 $(\alpha > \alpha_{ceil})$ には途中で直線探索を終了し、$J(w)$ の傾きが 0 の場合は直線探索を 1 回行い、1 ステップ目の点を最小点とする。本研究では、それぞれ$\alpha_0 = 10^{-5}$、$\alpha_{ceil} = 10^{20}$、$\alpha_{solv} = 10^{-3}$とした。以上、図 3のように V 型 3 点を見いだすことによって $J(w)$ の最小点を探索する。

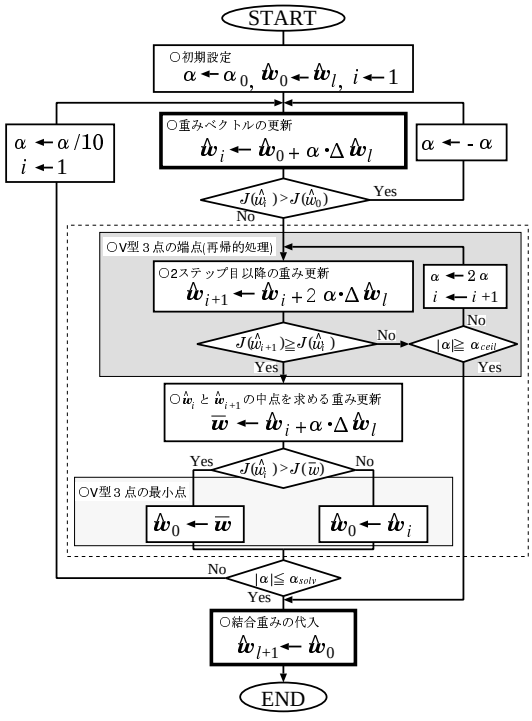
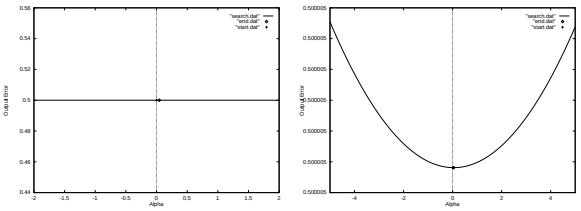



Fig. 3　等間隔 V 型 3 点探索法

**直線探索をする曲面**　共役勾配法に基づく学習アルゴリズムにおける直線探索では実際に図 4〜図 6のような誤差曲面 (断面) を直線探索している。



(a) 最小値付近の拡大図　　(b) 全体図

Fig. 4　探索の初期点と探索の終了点の誤差が同じ誤差曲面
学習回数 1300 回目; 初期重み 1400 試行目

図 4(a) を見ると誤差曲面は平坦であるが、図 4(b) を見ると、探索の初期点が既に誤差曲面の最小値であることが分かる。

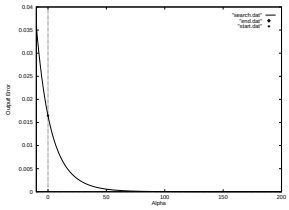
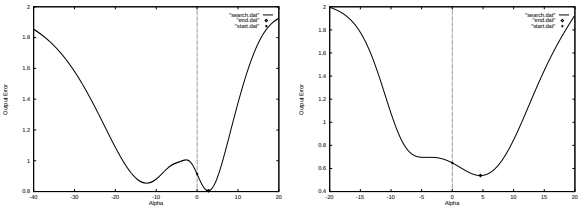


Fig. 5　最小点が存在しない誤差曲面：学習回数 59 回目; 初期重み 15 試行目

また、誤差が減少し続ける図 5のような誤差曲面の場合は最小値は存在しない。



学習回数 34 回目; 初期重み 7 試行目　　学習回数 29 回目; 初期重み 45 試行目

Fig. 6　多峰性の誤差曲面の断面

さらに、誤差曲面には図 4のような単峰性の誤差曲面だけでなく、図 6のような多峰性の誤差曲面も存在する。多峰性の誤差曲面の場合、探索の初期点によっては到達できる極小点が異なることがわかる。

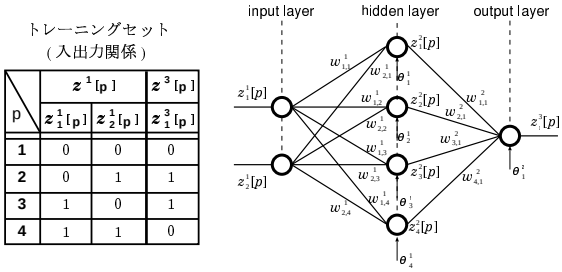
# 3.　重み空間における解析

## 3.1　排他的論理和問題における学習特性

本研究では、バイナリ問題である線形分離不可能な排他的論理和 (XOR) の学習問題を取り上げる。XOR 学習問題とは、入力パターン $(z_1^1[p], z_2^1[p])$ に対する所望な出力パターン $z_1^3[p]$ が、

$$\{(0,0),0\},\ \{(0,1),1\},\ \{(1,0),1\},\ \{(1,1),0\}$$

で与えられる入出力関係の組を学習する問題である。図 7に入出力関係の表とネットワークの構成を示す。




| p | $z^1[p]$ | | $z^3[p]$ |
|---|---|---|---|
| | $z_1^1[p]$ | $z_2^1[p]$ | $z_1^3[p]$ |
| 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 1 |
| 3 | 1 | 0 | 1 |
| 4 | 1 | 1 | 0 |

Fig. 7　XOR 問題の入出力関係とネットワーク

この 2-4-1 ネットワークにおける重みベクトル $\boldsymbol{w}$を次のように定義する。

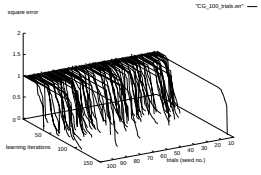
$$\boldsymbol{w} = [\theta_1^1, w_{1,1}^1, w_{2,1}^1, \theta_2^1, w_{1,2}^1, w_{2,2}^1, \theta_3^1, w_{1,3}^1, w_{2,3}^1,$$
$$\theta_4^1, w_{1,4}^1, w_{2,4}^1, \theta_1^2, w_{1,1}^2, w_{2,1}^2, w_{3,1}^2, w_{4,1}^2]^T \in \mathcal{R}^{N_w}$$

重みベクトル$\boldsymbol{w}$の次元の数はネットワークにおける結合重みとしきい値の数と等しく、次式で与えられる。

$$N_w = (N_1 + 1) \times N_2 + (N_2 + 1) \times N_3$$

この例では重みベクトルは 17 次元となる。このネットワークに対して、乱数により $(-0.1 \sim 0.1)$ の範囲で初期化された $10^4$個の異なる重みベクトル $\boldsymbol{w}^{[i]}(i = 1, 2, 3, \cdots, 10^4)$ を用いて、BP、$H_2$学習、$H_\infty$学習、および共役勾配法に基づく学習アルゴリズムを評価する。

図 8に共役勾配法の初期 100 回の試行に対する学習曲線、学習結果を示す。



学習曲線 (100 試行分)

学習結果

| $\beta$ | 0.4 | 0.0 |
|---|---|---|
| 最大値 | 2792 | 5702 |
| 最小値 | 9 | 3 |
| 平均値 | 58.6 | 228.7 |
| 分散 | 1482.5 | 77871.5 |
| 正規化分散 | 0.432 | 1.489 |
| Trap 数 | 9 | 5 |

Fig. 8　共役勾配法の学習結果 ($\alpha_0 = 10^{-5}$)

**他の学習アルゴリズムとの比較**　表 1に各学習アルゴリズムの学習結果、図 8の共役勾配法の結果と比較する。

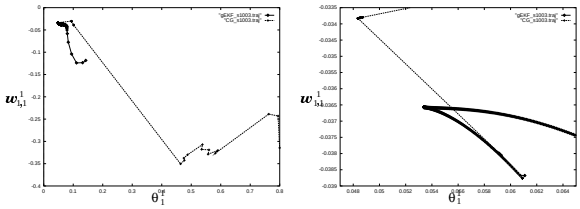
Table 1　他の学習アルゴリズムにおける学習結果

| アルゴリズム | BP ( $\eta$=0.8, $\beta$=0.8 ) | $H_2$学習 | $H_\infty$学習 ($\gamma_f$=1.7) |
|---|---|---|---|
| 最大値 | 9044 | 389 | 40 |
| 最小値 | 57 | 19 | 16 |
| 平均値 | 121.0 | 47.1 | 23.9 |
| 分散 | 10751.8 | 541.3 | 10.1 |
| 正規化分散 | 0.734 | 0.244 | 0.018 |
| Trap 数 | 0 | 0 | 0 |

以上より、共役勾配法に基づく学習アルゴリズムは他の学習アルゴリズムと比較すると、Trap 数が多いことが分かる。また、学習回数の最小値が他の学習アルゴリズムより小さいものの、学習回数のばらつきは BP と比較すると少ないが、$H_2$学習、$H_\infty$学習と比べると非常に多い。このことから、共役勾配法に基づく学習アルゴリズムは初期重みの変化に対するロバスト性に乏しいことが分かった。次章では共役勾配法の挙動について詳細に考察する。

## 4.　誤差曲面における学習過程

$H_2$学習と共役勾配法が、同じ初期重みベクトルから学習を開始し、その後どのような軌跡をとるか比較をすることによってその本質的な違いを明らかにする。そのため、$H_2$学習において最も学習回数の多かった 1003 試行目に着目し、2 次元投影面上の学習の軌跡を追跡した結果を図 9に示した。



(a) 全体図　　(b) 拡大図

Fig. 9　$H_2$学習における最悪ケース (1003) 試行目の学習軌跡 ($\theta_1^1$-$w_{1,1}^1$平面投影図)
学習回数：$H_2$学習 389 回, 共役勾配法 90 回

図 9(b) から $H_2$学習と共役勾配法に基づく学習アルゴリズムによる学習軌跡は途中までほぼ似たような軌跡である。しかし、共役勾配法に基づく学習アルゴリズムは途中から $H_2$学習とは大きく異なる軌跡となる。以降、この 1003 試行目の学習過程について解析を行う。

**学習曲線の比較**　共役勾配法に基づく学習アルゴリズムにおける学習毎の出力誤差の推移を表した学習曲線を図 10に示す。この図から共役勾配法に基づく学習アルゴリズムは最初に誤差がわずかに減少し、しばらく誤差が一定となった後、一気に減少する。

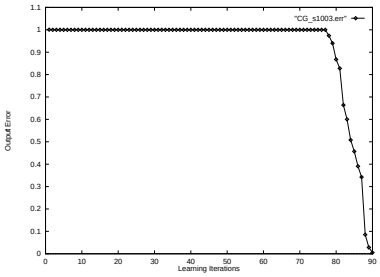



Fig. 10　1003 試行目の共役勾配法の学習曲線 ($\alpha_0 = 10^{-5}, \beta = 0.4$)

**学習毎の更新量の比較**　1003 試行において更新毎の重みベクトルの差分ノルム $d_l$ を

$$d_l = \sqrt{(\hat{\boldsymbol{w}}_l - \hat{\boldsymbol{w}}_{l-1})^T(\hat{\boldsymbol{w}}_l - \hat{\boldsymbol{w}}_{l-1})} \qquad (1)$$

と定義し、前学習からどれだけの量が変化したかを求めた。ここで、$\hat{\boldsymbol{w}}_l$ は $l$ サイクル目の学習で得られた重みベクトルであり、$\hat{\boldsymbol{w}}_0$ は重みベクトルの初期値を表す。図 11から共役勾配法に基づく学習アルゴリズムは学習開始直後から 76 回目まで極めて 0 に近い値で更新しているが、77 回目からは大きい値で更新している。

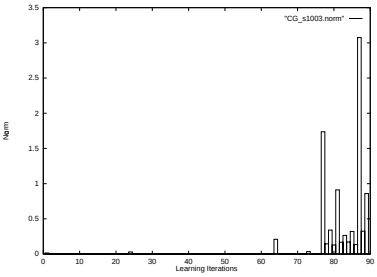



Fig. 11　1003 試行目の共役勾配法における学習毎の更新量 ($\alpha_0 = 10^{-5}, \beta = 0.4$)
学習回数：90 回

**誤差曲面の最急降下方向と重み更新方向**　$l$ 回目の学習サイクルにおける出力誤差の 2 乗和を $J[l] = \sum_{k=1+(l-1)N_p}^{lN_p} (y_k - h_k(\hat{\boldsymbol{w}}_{l-1}))^2$ とする。ただし、$N_p$ はパターン数を表す。この $J[l]$ を重みベクトル $\boldsymbol{w}$ で偏微分し、符号を反転すれば、以下のような Batch 型更新における誤差曲面の最急降下方向が得られる。

$$-\frac{\partial J[l]}{\partial \boldsymbol{w}} = 2 \sum_{k=1+(l-1)N_p}^{lN_p} \boldsymbol{H}_k^T (y_k - h_k(\hat{\boldsymbol{w}}_{l-1})) \qquad (2)$$

この誤差曲面の最急降下方向 $-\frac{\partial J[l]}{\partial \boldsymbol{w}}$ と、学習毎の重みの更新ベクトル $\Delta\hat{\boldsymbol{w}}_l$ との角度の余弦を、次式のように内積 $< \cdot, \cdot >$ を用いて求める。ただし、$l = 1, 2, \cdots, l_{end}$ であり、$l_{end}$ は学習終了時の学習回数を表す。また、$\hat{\boldsymbol{w}}_0$ は初期重みである。

$$\cos\theta[l] = \frac{\left\langle -\frac{\partial J[l]}{\partial \boldsymbol{w}}, \Delta\hat{\boldsymbol{w}}_l \right\rangle}{\| -\frac{\partial J[l]}{\partial \boldsymbol{w}} \| \cdot \| \Delta\hat{\boldsymbol{w}}_l \|} \qquad (3)$$
$$\Delta\hat{\boldsymbol{w}}_l = \hat{\boldsymbol{w}}_l - \hat{\boldsymbol{w}}_{l-1}$$

共役勾配法に基づく学習アルゴリズムを用いたときのこの更新毎の余弦を図 12に示す。

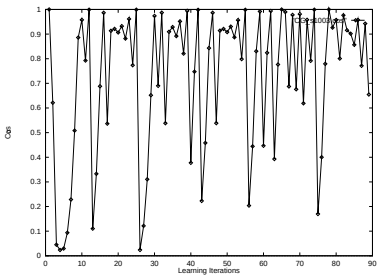
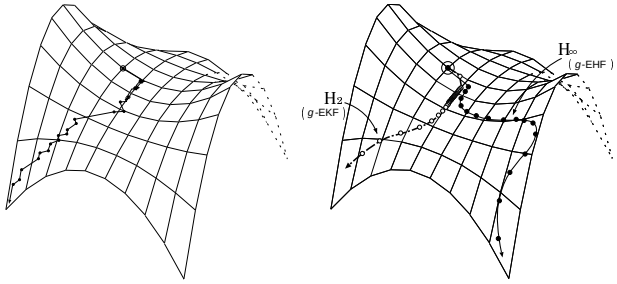



Fig. 12　学習毎の誤差曲面の最急降下方向と重み更新方向との余弦 ($\alpha_0 = 10^{-5}, \beta = 0.4$)
共役勾配法 学習回数 90 回 , 初期重み 1003 試行目

この図より、共役勾配法に基づく学習アルゴリズムでの重みの更新方向は誤差曲面の最急降下方向に対して、沿っていたり、直交していたりと非常に不安定な挙動をとっていることが分かる。以上の結果から $H_2$ 学習、$H_\infty$ 学習、共役勾配法に基づく学習アルゴリズムが所望の解を得るまでの誤差曲面上の挙動を 3 次元空間において模擬的に表したものを図 13に示す。




(a) 共役勾配法
($\alpha_0 = 10^{-5}, \beta = 0.4$)
学習回数：90 回

(b)$H_2$ 学習
学習回数：389 回
$H_\infty$ 学習
学習回数：27 回

Fig. 13　最悪ケースにおける共役勾配法,$H_2$ 学習,$H_\infty$ 学習の誤差曲面上の挙動; 初期重み 1003 試行目

図 13から、共役勾配法に基づく学習アルゴリズムは誤差曲面の鞍点上で振動していることが分かる。このことから、図 13のように鞍点の尾根に非常に近いところに初期重みがあった場合、更新の過程で鞍点に因われてしまい、鞍点から脱出するのが困難になる。

## 5. 結論

共役勾配法学習アルゴリズムの重み空間および誤差曲面上での挙動について詳しく解析した結果、鞍点上で細かく振動していることが分かった。そのため学習回数が非常に多くなる試行もあることから、初期重みの変化に対するロバスト性に乏しく、有効な学習法であるとは言い難いと思われる。直線探索をしている誤差曲面の断面を見ると、直線探索できない曲面や、更新方向によっては真の最小点を探索できない曲面もあることが分かった。このことからニューラルネットワークの学習法において、直線探索を用いた学習法は好ましくないと思われる。

## 参考文献


1) 西山 清: “最適フィルタリング”, 培風館 (2001).

2) 西山 清 , 落宰 公志: “大域的準最適 $H_\infty$学習の挙動”, 信学技報 , NC2002-144, pp.63-70 (2003).

3) 西山 清 , 落宰 公志 , 渡辺 充範: “$H_\infty$学習の何が新しいか? -共役勾配法、準ニュートン法との比較を中心に-”, 信学技報 , NC2003, to appear, 2004.

4) 嘉納 秀明: システムの最適理論と最適化, コロナ社 (1986).

5) 電気学会: 学習とそのアルゴリズム -ニューラルネットワーク ・遺伝アルゴリズム ・強化学習-, 森北出版 (2002).