バックプロパゲーション法学習過程における効率評価尺度の検討

西村治彦† 小山宣樹†

階層型ニューラルネットワークにおけるバックプロパゲーション法学習則は，その応用上の可能性から様々な研究が展開されてきた．ただ，これまでの研究では，ネットワークの学習性能評価に際して最終的な誤差関数値を重視する，すなわち学習曲線における最初と最後の落差に注目する傾向が強かった．我々は，性能や能力なる量は学習時のプロセス全体を考慮して判断されるべきであるという観点から，学習過程の効率評価尺度を新たに導入する．これによって従来，パターンとして目で捉えていた学習曲線の経時的ふるまいを数値として定量化することができた．論文では，実際の文字学習実験でのデータをもとにこの評価指数（L値）の性質と有効性について検討している．その結果，学習負荷による学習過程への影響の違いを評価できること，値が小さいほど認識時に対ノイズ・センシティビティが良い傾向をもつことなど，学習終了時（打ち切り時）の誤差関数値には見られない独自性が確認できた．最近では，学習課題に対して自己適応的にシステム変更してゆく方法が注目され始めているが，そこでの適合度尺度の一つとしてもL値は活用できそうである．

A Consideration on Measure of Efficiency of the Back-Propagation Learning Process

Haruhiko Nishimura† and Nobuki Koyama†

In this paper, we propose a new measure of efficiency of the back-propagation learning process, called L value, which can evaluate numerically the whole behavior of a learning curve. Its characteristics are investigated experimentally based on alphabet characters learning simulations. Some properties are found concerning the relations to task's loads and to noise (defect) levels. L value is more favorable as a performance measure than the ordinary error function and is expected to be a fitness function in auto-adaptive neural network schemes.

1. はじめに

階層型ニューラルネットワークの学習アルゴリズムとしてバックプロパゲーション(BP)法学習則[1),2)]が提出されて以来，文字認識，音声認識，画像処理をはじめ様々な分野で応用されている．その中で，勾配法に由来するローカルミニマムへの捕捉や学習収束速度の鈍化などの問題が指摘され，その改善策が研究されてきた[3),4)]．このほかに実用上のネットワークの学習能力を左右する重要な問題として，応用事例ごとにネットワークの構造および種々の学習パラメータをどのように決定するかがある．具体的には，中間層数，中間層のユニット数，学習係数，ユニットの入出力特性などの決定方法である．しかし，通常は学習前にはっきりとした指針はないので，適当なパラメータの組み合せを試行錯誤的に試しているのが現状である．最近では，むしろ事前情報に期待せず，自己適応的にシステム変更を行っていく方法が注目され始めている[5),6)]．

ところで，これまでの研究では，学習評価尺度である誤差関数値（またはそれに付加項をつけたもの）が予め設定されたしきい点に到達するまでの所要時間が短いほど学習能力が高い，または，一定の時間内にそれが到達した値が小さいほど学習性能が良いとされがちであった．これは，誤差関数の経時的変化である学習曲線でいえば，最初と最後の落差を重視し，その間の経路の違いを軽視するものであり，ネットワークの学習性能を学習終了（打ち切り）時の最終誤差関数値で評価することにつながる．しかし，性能や能力なる量は学習の最終時の状況のみから評価できるものではなく学習時のプロセス全体を考慮して判断されるべきものである．すなわち，学習過程全体を通しての学習効率によって計られるべきものであり，それを反映する新たな評価尺度が必要となる．

† 兵庫教育大学情報科学研究室
Department of Information Science, Hyogo University of Teacher Education

このような観点から本論文では，ネットワークにおける学習過程全体の効率評価のための尺度として，新たに学習効率評価指数を定義し，その性質と有効性について実験的に検討する．この評価量は，従来パターンとして目で把えていた学習曲線のふるまいを数値として定量化するもの[7]で，今後の自己適応的なシステム変更モデルにおける適合度尺度の一つとしても活用可能である．

以下，第2章では，バックプロパゲーション法学習則について簡単に触れるとともに，学習過程全体の効率を記述する学習効率評価指数を新たに定義，導入する．第3章では実際の文字学習実験を通してこの指数を具体的に求め，そこでのデータをもとに第4章では最終誤差関数値との相関，学習負荷との相関，学習後認識時のノイズ影響との関係からその性質と有効性について検討する．最後に第5章で，本論文での結論と今後の課題について述べる．

2. BP 法学習則と学習効率評価指数

階層型ニューラルネットワークは入力層，出力層およびいくつかの中間層からなり，入力から出力方向への層間結合はあるが層内結合はない．中間層，出力層の各ユニットの入出力関係は，

$$y_i=f(\sum_j w_{ij}x_j) \tag{1}$$

で与えられる．ここで w_{ij} はユニット j から次層のユニット i への結合荷重，$\{x_j\}$ はユニット j からユニット i への入力値である．しきい値はバイアスユニット（$j=0$）からの入力（$x_0=1$）による結合荷重 w_{i0} の効果として解釈できる．出力関数 f としては有界で単調増加のシグモイド関数

$$f(s)=\frac{1}{2}\left(1+\tanh\left(\frac{s}{\kappa}\right)\right) \tag{2}$$

をとる．κ はシグモイド関数の傾き係数で，$\kappa\to 0$ の極限で $f(s)$ は階段関数に一致する．

ネットワークの入力層にパターン p を提示した際の出力層ユニット i の出力を o_{pi}，教師データによる目標出力を d_{pi} とするとき，誤差関数は，

$$E=\sum_p E^{(p)}=\sum_p\sum_i (o_i^{(p)}-d_i^{(p)})^2 \tag{3}$$

で定義される．BP 法学習則はこの E を最小とする結合荷重およびしきい値を最急降下法を用いて逐次近似的修正によって求めてゆこうとするものである．その修正量は，

$$\Delta w_{ij}=\eta\delta_i x_j \tag{4}$$

によって与えられる．ここで，η は学習係数である．δ_i は，ユニット i が出力層にあるか，中間層にあるかで異なる．今，ユニット i への入力の総和を S_i とすると，出力層の場合には，

$$\delta_i\equiv\delta_{2i}=(d_i-o_i)f'(S_i) \tag{5}$$

中間層の場合には，

$$\delta_i\equiv\delta_{1i}=(\sum_l \delta_{2l}w_{li})f'(S_i) \tag{6}$$

となる．ただし，$f'(s)$ は関数 f の微分である．

誤差関数 E の値は，ネットワークが正しい答にどのくらい近づいているかを表す学習の定量的尺度としての意味を持つ．学習回数に対する E 値の経時変化全体は通常，学習曲線と呼ばれている．つまり，学習曲線は学習が収束に向かう履歴を記述するものである．この学習曲線の違いを数値化することでネットワーク全体の学習効率の定量的評価が可能になる．

そのためにまず，E 値の学習回数に対する経時加重平均

$$F_\alpha=\frac{\sum_{t=1}^{t} E_t\sqrt[\alpha]{t}}{\sum_{t=1}^{t}\sqrt[\alpha]{t}} \tag{7}$$

と E 値の変化率の経時加重平均

$$G_\beta=\frac{\sum_{t=1}^{t}\{\log_{10}(E_{t-1}/E_t)\}(1/\sqrt[\beta]{t})}{\sum_{t=1}^{t}(1/\sqrt[\beta]{t})} \tag{8}$$

を導入する．ここで，α，β は F，G それぞれの累乗根のパラメータであり，t は学習時刻（学習の進行回数）である．一般に効率の良い学習過程とは，t の早い時期に E_t が学習毎に大きく低下し，その後もゆるやかながら順調に減少する過程をさす．F_α, G_β で言えば，F_α が小さく G_β が大きいほどこの傾向が強いことになる．そこで我々は，両者の比

$$L=\frac{F}{G} \tag{9}$$

なる量を考え，学習過程における学習効率の定量的評価の尺度とする．つまり，この L の値が小さいほど学習の際の効率は良いということになる．これ以後，この L の値のことを我々は学習効率評価指数（L 値）と呼ぶ．

3. 学習実験

学習タスクとして，ここではアルファベット文字学習を選択した．入力パターンは，**図1**に示すようにそれぞれの文字を16行16列のドット表現とし，各画素の■，□状態を256個の入力層ユニットの1，0状態に対応させて構成した．また，目標出力である教師パターンは，文字種に対応する1箇所のみが1で，他は

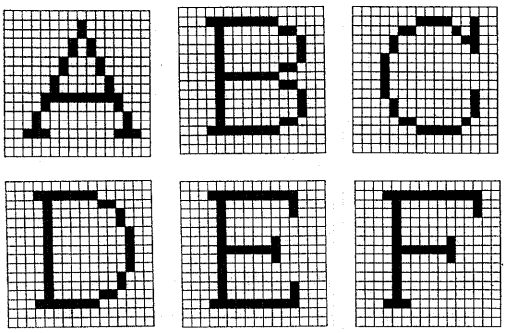


図 1 文字学習実験用入力パターンデータ例
Fig. 1 Examples of training patterns used in our experiment.

0であるような直交パターンとした．例えば，Aの場合の教師パターンは $d^{(A)}=(1,0,0,\cdots,0)$，Bの場合は $d^{(B)}=(0,1,0,\cdots,0)$ 等となる．このとき，出力層のユニット数は学習文字数 (N_p) 分必要になる．

ネットワークは3層構造とし，データ取得に際しては，中間層のユニット数は 20，シグモイド関数の傾き係数は $\kappa=0.75$，学習係数は $\eta=0.5$ とした．結合荷重としきい値の初期値は一様乱数 ($\leqq|0.2|$) で与えることとした．以上の設定の下，1パターン提示毎に結合荷重としきい値を修正する逐次修正法によって学習を実行した．

データとしては，対象のアルファベット N_p 個を100回 ($f=100$) 学習させ，学習毎の誤差関数 E_t の値を取得し，(7)から(9)式を用いて L 値を計算した．ただし，ここでは，N_p 個一巡で学習1回と呼んでおり，結合荷重としきい値の修正回数でいえば $100N_p$ 回ということになる．

結合荷重としきい値の初期値が異なる100例に対する学習実験結果から，$F-G$ の負の相関が確認できた．$N_p=26$ の場合にきわめて収束性の悪い4例を除いた96例について，$F-G$ 相関 ($\alpha=2$, $\beta=3$) を図示したのが**図2**である．G→大，F→小のとき，学習効率→良という負の相関がよく現れている．

さらにこのうちの4例について，その E 値の経時変化をグラフ化したのが**図3**である．値が小さくなってもその変化を追えるよう，図の縦軸は対数表示をとっている．図中の(1)～(4)の曲線のそれぞれの履歴には，かなりの違いが存在しているのがわかる．これらに対して L 値を求めてみると，それぞれ，(1) 39.22，(2)55.04，(3)71.23，(4)94.19 となる．この結果からも，各学習過程にみられる効率の良し悪しが，L値によって定量的に正しく評価されていることがわかる．

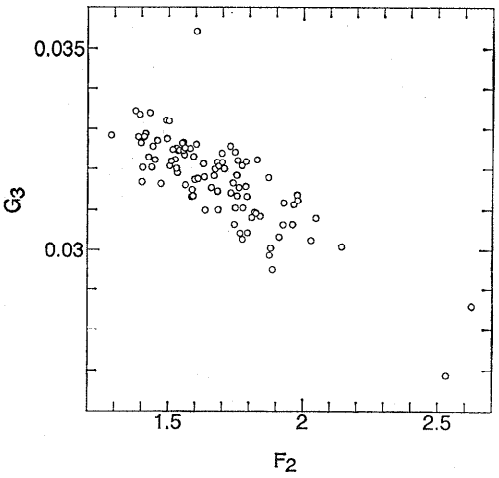


図 2 文字（A～Z）学習における F_2-G_3 相関
(100回学習データ96例による)
Fig. 2 F_2 versus G_3 calculated by eqs. (7) and (8) for 96 trials. (Number of training patterns: $N_p=26$, Iteration number: $f=100$.)

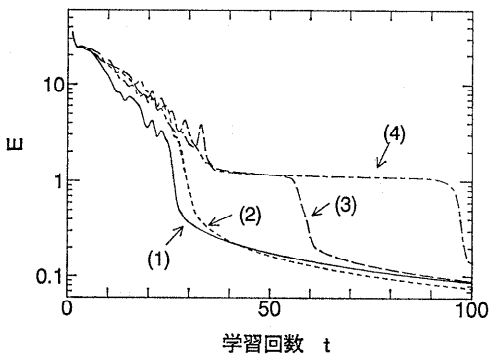


図 3 文字（A～Z）学習における学習曲線例
Fig. 3 Learning behaviors for typical 4 examples.

4. L 値の性質とその有効性

ここでは第3章の文字学習実験のデータをもとに，学習打ち切り時点での誤差関数値（最終誤差関数値）との比較を通して，L値の性質と有効性について評価する．

4.1 最終誤差関数値（E_f 値）との相関

結合荷重としきい値の初期値を変えて行った文字学習実験100例に対して，その $L-E_f$ 相関（今，学習回数は100なので $f=100$）を図示したのが**図4**であ

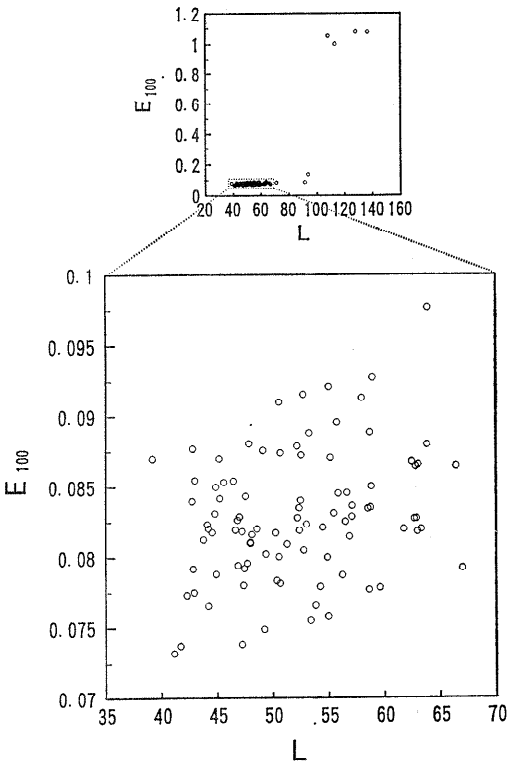


図 4　文字（A～Z）学習における $L-E_f$ 相関
(100 回学習データ，100 例による)
Fig. 4　L versus E_f for all 100 trials.
($N_p=26$, $f=100$)

る．100 例全てに対する大局図（上図）では，収束性のきわめて悪い 4 例に対しては，両者ともはずれ値を与えているが，それ以外では，E_{100} はきわだって局在しており，学習過程全体の学習効率の尺度性は弱い．また，図中の [ ] 部分（両者の低域）を拡大した下図においても，両者の間に全体として弱い相関は認められるものの，互いの順位相関などはみられない．このことから，従来の E_f 値とは違う，学習過程の効率評価尺度としての L 値の固有性が確認できた．

4.2　学習負荷との相関

一般に，ニューラルネットワークにおける学習過程は学習負荷（その時の学習課題の種類や量）に依存する．ここでは，学習負荷の違いが E_f 値と L 値にどのように反映されるかを調べてみる．実験では，文字学習の際の学習パターン総数 N_p を違えることで負荷の違いを設定した．ただ，第 3 章で設定したネットワークでは，出力層ユニット数 N_0 は $N_0=N_p$ の関係で変化するので，負荷の違いの影響だけを見るため

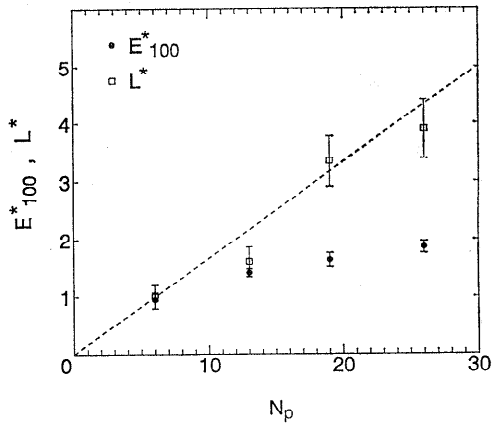


図 5　E^*_f, L^* の学習パターン数依存性（縦軸は $N_p=6$ の場合を 1 とする規格化表示．各々 94 例平均）
Fig. 5　Dependence of E^*_f and L^* on the number of training patterns. (Values are normalized by the value in $N_p=6$ case and are averaged from 94 trials.)

には(3)式の誤差関数値そのままではなく，それを N_0 で割った 1 出力ユニットあたりの誤差関数値

$$E^*=\frac{E}{N_o} \tag{10}$$

を考える必要がある．このとき(7)式の F_α は

$$F^*_\alpha=\frac{F_\alpha}{N_o}$$

となるが，(8)式の G_β はそのままなので，1 出力ユニットあたりの L 値は，E 値同様，

$$L^*=\frac{L}{N_o} \tag{11}$$

で与えられることがわかる．

具体的には，$N_p=6, 13, 19, 26$ のそれぞれの場合について，結合荷重としきい値の初期値が異なる 100 例を用意し，そのデータを取得した．それをもとに N_p と E^*_f($f=100$) 及び L^* の平均値との関係をグラフにしたのが**図 5** である．E^*_f, L^* の傾向を同じスケールで比較するために，図の縦軸は $N_p=6$ のときのそれぞれの値で割った規格化表示となっている．なお，平均に際しては，各パターン数において学習収束性のきわめて悪い．すなわち，L, E_{100} の極端なはずれ値をもつ例は除外した（$N_p=6$, 13, 19, 26 の場合にそれぞれ 6 例ずつ）．誤差棒は平均値±標準偏差を示すものである．このグラフから，N_p の増加に対して，L^* はほぼ比例して増加しているが，E^*_{100} は増加が鈍く，一定化の傾向が見受けられる．このことは，学習効率評価指数（L値）は，ニューラルネットワークへ

の学習負荷の違いを定量的に評価する尺度という観点からも有効であると考えられる．

4.3 学習後認識時のノイズ影響との関係

学習終了後のニューラルネットワークに，学習時使用した文字パターンにノイズ（欠陥）を付加したパターンを認識させると，そのノイズ（欠陥）の影響のため，認識時に得られる誤差関数値は学習終了（打ち切り）時の最終誤差関数値からずれてくる．ここでは，そのずれの度合いと学習効率評価指数（L値）の関係を，E_f 値との比較を通して調べることにする．

実験では，認識用のノイズ（欠陥）ありパターンとして，パターンAの■部分から1，2，3，4，5個抜き取った5パターン（$n=1\sim5$）を用意した．次に，学習終了（打ち切り）時の結合荷重としきい値をもつニューラルネットに，これらのパターンを入力し，そのとき得られる誤差関数値 $E^{(A)}$ が，ノイズのないとき（$n=0$）の $E^{(A)}$ と比べてどれだけずれたかを評価するため，

$$D=\frac{|E^{(A)}_{(n)}-E^{(A)}_{(0)}|}{E^{(A)}_{(0)}} \qquad (12)$$

なる量を用いることにする．ここで，$E^{(A)}=\sum_i(d_i^{(A)}-o_i^{(A)})^2$，$(n)$ の n はノイズ（欠陥）数である．

まず初めに，このズレ率 D のノイズに対する性質を見ておくことにする．学習パターン数の違う四つの場合について，ノイズ数 n と D の関係をグラフ化したのが**図6**である．値は先の 94 例の平均値である．

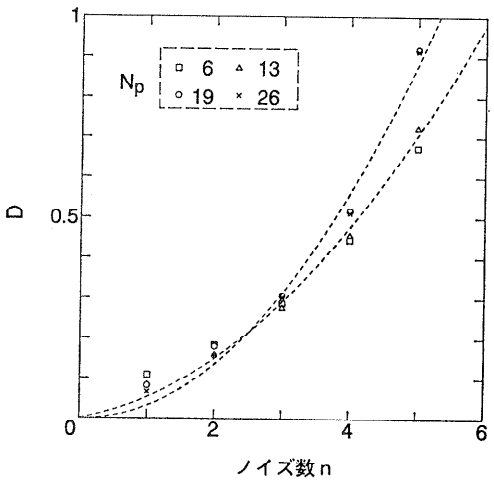


図6 ノイズ数 n とズレ率 D の関係（学習パターン数 N_p が違う四つの場合．D の値は 94 例平均）
Fig. 6 Distortion rate D as a function of the number of defects n for N_p=6, 13, 19, 26. (Each value of D is averaged from 94 trials.)

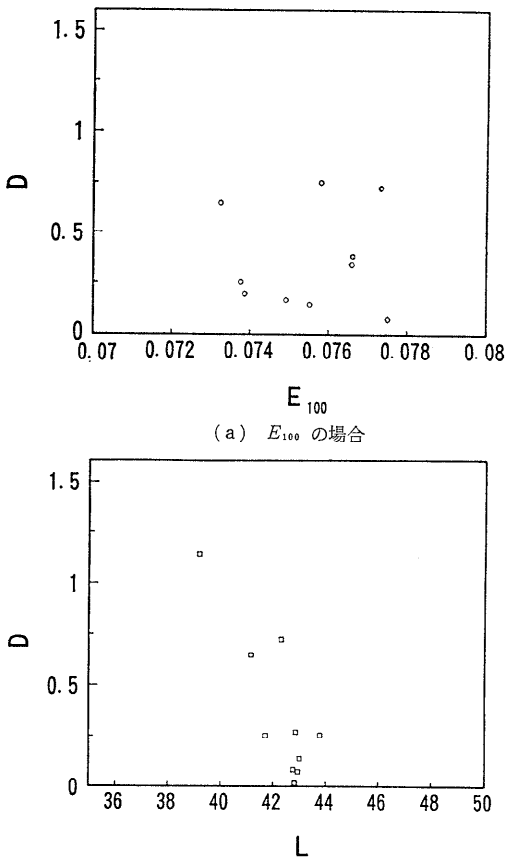


図7 E_f, L それぞれの小さい側から 10 番目までの 10 例における D 値依存（N_p=26, n=3 の場合）
Fig. 7 Dependence of D on $E_f(L)$ evaluated for 10 events from the lowest to the lower tenth in $E_f(L)$. (N_p=26, n=3.) (a) E_{100} case, (b) L case.

どの N_p の場合も，n とともに D の値はほぼ $O(n^2)$ で増加しているが，その増加傾向は N_p=6, 13 に比べて N_p=19, 26 の方が急なものとなっている．これは，学習パターン数が多いほど各パターン間の切り分け学習（特徴抽出）が難しく，その結果，学習後のネットワークはノイズに対して敏感になるためと解釈される．

さて，N_p=26 の場合の E_{100}，L それぞれについて低域側の（つまり，値として良の）10 サンプルに着目し，n=3 のときの D 値分布を図示したのが**図7**である．（a），（b）双方の分布にははっきりした違いが存在する．$L-D$ の間に，L が良い（小さい）とき D が大きくなる負の相関傾向が見られ，E_{100}，L の低域においては，E_{100} よりも L の方がノイズに対してセ

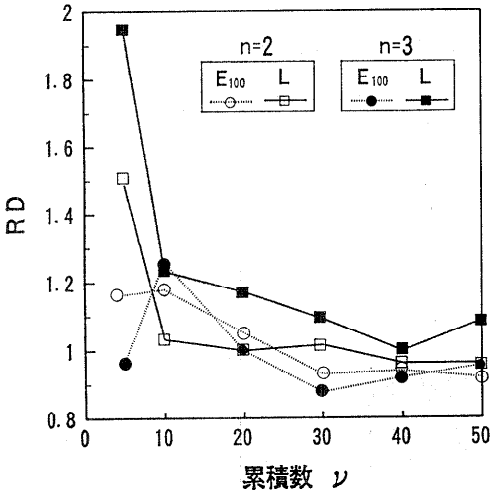


図 8　E_{100}, L に対する RD と累積数 ν の関係（ノイズ数 $n=2,3$ の場合）

Fig. 8　Relation between RD and the cumulative number ν for E_{100} and L. (Number of defects: $n=2,3$.)

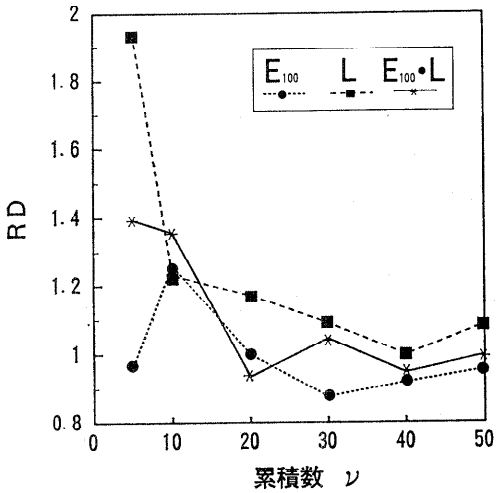


図 9　$E_{100}\cdot L$ に対する RD と累積数 ν の関係（ノイズ数 $n=3$ の場合．比較のため E_{100}, L に対するグラフも表示．）

Fig. 9　Relation between RD and the cumulative number ν for $E_{100}\cdot L$ compared with the cases for E_{100} and L. (Number of defects: $n=3$.)

ンシティブであることがわかる．

さらに，D を用いて次の RD を定義し，対ノイズ性を調べてみる．

$$RD=\frac{\bar{D}_{(\nu)}}{\bar{D}_{(total)}} \tag{13}$$

この RD は，対象とする全学習例から E_{100}, L 値の小さいもの順に ν 個選び出し，その ν 個が与える D の平均値 $\bar{D}_{(\nu)}$ を全学習例の D の平均値 $\bar{D}_{(total)}$ で割ったものである．ノイズ数 $n=2,3$ の場合について，E_{100}, L それぞれでの累積数 ν に対する RD の変化を**図 8** に示した．E_{100} の場合に比べて L の場合，RD の値が ν の小さい側で大きく跳ね上がっている．このことからも，E_{100}, L 両者の対ノイズ性の違いがよくわかる．

ここでは L と E_f の違いについて注目してきたが，対ノイズ性を評価する尺度としては L と E_f を組み合わせて利用することも可能である．今，一例として $E_f\cdot L$（E_f と L との積）を取り上げてみよう．すると，$E_f\cdot L$ の小さいものから ν 個に対する RD とその ν との関係は**図 9** のようになる．$\nu=5$ での RD の値が E_{100}, L それぞれの場合の中間程度になっているのがよくわかる．このことは，E_{100} と L をうまく用いることにより，必要な RD をもつ学習例の集団を選び出せる可能性を示している．

5. おわりに

学習効率評価指数（L値）の導入によって，ネットワークの種々の学習過程の違いを数値的に定量化することが可能となった．この L 値は，従来の評価尺度である誤差関数値を基に構成されているが，学習終了時（打ち切り時）の最終誤差関数値とは違う独自性をもつ．このことは，学習負荷による学習過程への影響の違いを評価できるという点や，値が小さいほど認識時に対ノイズ・センシティビティが良い傾向をもつという点に現れている．さらに，この学習効率評価指数と最終誤差関数値を組み合わせることで，学習効率が良く，ノイズに対して好みのロバストネスを持つ学習例を選択できる可能性も示すことができた．

ここでは，結果の信頼性と一般性を失わないために，結合荷重としきい値の初期値については 100 種類用意し，統計的な立場から議論を進めてきた．学習パターンとしてはアルファベットの場合の結果しか示さなかったが，記号（例えば○□△＋×…）等の別のパターン内容の場合にも，同様の傾向の結果を実験的に確認している．また，学習回数を $f=100$ から $f=200$ へ増やした場合の実験でも，L 値が全体的に小さくなるだけで，結果の傾向に変わりはなかった．

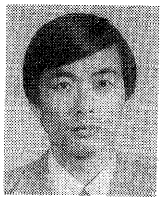
学習効率評価指数の活用という観点からは，結合荷重の初期値，中間層のユニット数，ユニットの入出力

特性，学習係数などの種々のパラメータがネットワークの学習性能に与える影響を L 値を用いて統一的に検討することが可能となる．各パラメータの L 値への影響力を様々な学習負荷に対して実験的に分析し，効果的に L 値を小さくする方策を探ることは今後の重要な課題の一つである．また，L 値と学習後認識時におけるノイズ・センシティビティの関係が明確になれば，多数の学習例から好みのロバストネスをもった学習例を選択できる可能性があり，現在応用上直面している BP 法学習則の汎化能力や過学習の問題への現実的な対応策の提供につながり得る．多数の学習例からの選別に際しては，生物の適応過程（進化過程）をルール化した遺伝的アルゴリズム[8]に着目し，L 値をその適合度指標とすることでロバストネス調節のシマテマティックな方法が構成できそうである．

参考文献

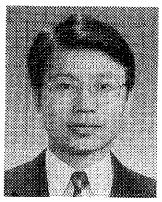
1) Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning Representations by Back-Propagating Errors, *Nature*, Vol. 323, No. 9, pp. 533-536 (1986).
2) Rumelhart, D. E., McClelland, J. L. and the PDP Research Group: *Parallel Distributed Processing*, Vol. 1, MIT Press (1986).
3) 倉田耕治，麻生英樹：神経回路網の理論的研究における最近の動向，電子情報通信学会論文誌，Vol. J 73-D-II, No. 8, pp. 1103-1110 (1990).
4) 喜多 一：ニューラルネットワークの汎化能力，システム/制御/情報，Vol. 36, No. 10, pp. 625-633 (1992).
5) 石川真澄：ネットワーク学習アルゴリズムの最近の話題，計測と制御，Vol. 30, No. 4, pp. 285-290 (1991).
6) Whitley, D., Starkweather, T. and Bogart, C.: Genetic Algorithms and Neural Networks: Optimizing Cconnections and Cconnectivity, *Parallel Computing*, Vol. 14, No. 3, pp. 347-361 (1990).
7) 西村治彦，小山宣樹：階層型ニューラルネット学習効率の入力パターン分解能依存性評価，第45回情報処理学会全国大会論文集（2），pp. 321-322 (1992).
8) Goldberg, D. E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley (1989).

（平成5年3月8日受付）
（平成5年7月8日採録）



西村 治彦（正会員）

1980年静岡大学理学部物理学科卒業．1985年神戸大学大学院自然科学研究科博士課程修了．1989年広島大学医学部（医療情報学）助手．1990年12月より兵庫教育大学（情報科学）助教授，現在に至る．学術博士．これまでファジー・データベース，ニューラルネットワーク，セルオートマトン，複雑系の科学等の研究に従事．神経回路学会，日本物理学会，システム制御情報学会等各会員．



小山 宣樹（正会員）

1953年生．1977年武蔵工業大学経営工学科卒業．現在，和歌山県立田辺工業高等学校教諭．1991～1993年兵庫教育大学大学院学校教育研究科情報科学研究室にてニューラルネットワークの研究に従事．修士（学校教育学）．現在，産業教育振興中央会の助成による産業教育改善に関する特別研究中．日本オペレーションズ・リサーチ学会，日本産業技術教育学会，環境情報科学センター各会員．