$=-E(v_1)$，$\hat{V}''(0)=E(v_1^2)$，可得(9)式。

而
$$\begin{aligned}
N(Z) &= E(Z^N)=E(Z^{X+N_1+N_2+\cdots+N_X})\\
&= \sum_{k=1}^{\infty} Z^k E\left[Z^{N_1+N_2+\cdots+N_{\xi(\sum_{i=1}^{k}v_i)}}\right]P\{X=k\}\\
&= \sum_{k=1}^{\infty} Z^k \int_0^{\infty} E\left[Z^{N_1+N_2+\cdots+N_{\xi(t)}}\right]\mathrm{d}P\left\{\sum_{i=1}^{k}v_i<t\right\}P\{X=k\}\\
&= \sum_{k=1}^{\infty} Z^k \int_0^{\infty}\sum_{n=0}^{\infty} E\left[Z^{N_1+N_2+\cdots+N_n}\right]P\{\xi(t)=n\}\mathrm{d}P\left\{\sum_{i=1}^{k}v_i<t\right\}P\{X=k\}\\
&= \sum_{k=1}^{\infty} Z^k \int_0^{\infty} \mathrm{e}^{-\lambda[1-N(Z)]}\mathrm{d}P\left\{\sum_{i=1}^{k}v_i<t\right\}P\{X=k\}\\
&= \sum_{k=1}^{\infty} Z^k E\left\{\mathrm{e}^{-\lambda[1-N(Z)]\sum_{i=1}^{k}v_i}\right\}P\{X=k\}\\
&= \sum_{k=1}^{\infty} Z^k \hat{V}[\lambda-\lambda N(Z)]\}^k P\{X=k\}\\
&= X\{Z\hat{V}[\lambda-\lambda N(Z)]\}
\end{aligned}$$

由上式易得(10)式。

2　批到达泊松近似过程的参数

现在我们来求批到达泊松近似过程的两个参数，一个是到达的平均间隔时间$\frac{1}{\lambda}$，另一个是每批到达的平均顾客(信息包)数 r. 为此，设 ATM 复用器的输入是由 N 个相互独立的 ON-OFF 源构成的迭加过程，每个 ON-OFF 源的活动期 ξ_i 均服从参数为 α 的指数分布，静止期 η_i 均服从参数为 β 的指数分布，即 $\xi_i\sim\Gamma(1,\alpha)$，$\eta_i\sim\Gamma(1,\beta)$，用 K 表示迭加过程中处于活动期的 ON-OFF 源的个数，显然，K 能取值为 $0,1,2,\cdots,N$，称 K 能取的值为这个迭加过程的相位。记 $R_0=\{0\}$，$R_1=\{1,2,\cdots,N\}$ 称 R_0,R_1 分别为迭加过程的状态 0 和状态 1.

对一个 ON-OFF 源来说，因为 $E(\xi_i)=\frac{1}{\alpha}$，$E(\eta_i)=\frac{1}{\beta}$，所以在任一时刻它处于活动期的概率 p 为
$$p=E(\xi_i)/E(\zeta_i)=\frac{\beta}{\alpha+\beta} \tag{17}$$
又因 N 个 ON-OFF 源独立同分布。所以在任一时刻，迭加过程中恰有 k 个处于活动期的概率为
$$P\{K=k\}=C_N^k p^k(1-p)^{N-k}\qquad k=0,1,\cdots,N,\text{即 } K\sim B(N,p) \tag{18}$$
显然，这个迭加过程就其相位来说，是一个生 - 灭过程。当它处于相位 i(即有 i 个 ON-OFF 源处于活动期)时，其生率为 $p_i=(N-i)\beta$，灭率为 $q_i=i\alpha$，$i=0,1,\cdots,N$. 所以其(相位)密度矩阵 Q 为
$$Q=\begin{bmatrix}
-N\beta & N\beta & 0 & 0 & 0 & \cdots & 0\\
\alpha & -[\alpha+(N-1)\beta] & (N-1)\beta & 0 & 0 & \cdots & 0\\
0 & 2\alpha & -[2\alpha+(N-2)\beta] & (N-2)\beta & 0 & \cdots & 0\\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\
0 & 0 & 0 & 0 & \cdots & N\alpha & -N\alpha
\end{bmatrix} \tag{19}$$

从而其平稳分布由 $\pi' Q = 0$ 与 $\sum_{i=0}^{N}\pi_i = 1$　（其中 $\pi' = (\pi_0, \pi_1, \cdots, \pi_N)$）

得
$$\pi_k = \frac{p_{k-1}p_{k-2}\cdots p_0}{q_k q_{k-1}\cdots q_1}\pi_0 = C_N^k\left(\frac{\beta}{\alpha+\beta}\right)^k\left(\frac{\alpha}{\alpha+\beta}\right)^{N-k}, k = 0,1,\cdots,N \tag{20}$$

此式与(18) 式相同。由(19) 式可立得相应跳跃链的概率转移矩阵 P：

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ \frac{\alpha}{\alpha+(N-1)\beta} & 0 & \frac{(N-1)\beta}{\alpha+(N-1)\beta} & 0 & 0 & \cdots & 0 \\ 0 & \frac{2\alpha}{2\alpha+(N-1)\beta} & 0 & \frac{(N-2)\beta}{2\alpha+(N-2)\beta} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \tag{21}$$

当迭加过程中恰有 i 个 ON-OFF 源处于活动期时，则其余 $(N-i)$ 个 ON-OFF 源必处于静止期。设　$X_i = \min(\xi_1, \xi_2, \cdots, \xi_i), Y_{N-i} = \min(\eta_1, \eta_2, \cdots, \eta_{N-i}), i = 0,1,\cdots,N$

其中，诸 ξ_i　i.i.d，诸 η_i　i.i.d.，　$X_0 = +\infty = Y_0$

设
$$Z_i = \min(X_i, Y_{N-i}), i = 0,1,\cdots,N$$

则从迭加过程开始处于相位 i 时起，经过时间 Z_i 后，相位必发生变化，或 $i \to i-1$，或 $i \to i+1$. 记 $i \to i-1$ 的概率为 μ_i，$i \to i+1$ 的概率为 λ_i. 则由指数分布的性质和全概率公式，易知

$$\begin{cases} \mu_i = P\{X_i < Y_{N-i}\} = \dfrac{i\alpha}{i\alpha+(N-i)\beta} \\ \lambda_i = P\{X_i > Y_{N-i}\} = \dfrac{(N-i)\beta}{i\alpha+(N-i)\beta} \end{cases} \quad i = 0,1,\cdots,N \tag{22}$$

记 $z_i = E(Z_i)$，则因 $Z_i \sim \Gamma(1, i\alpha + (N-i)\beta)$，故有

$$z_i = \frac{1}{i\alpha+(N-i)\beta}, \qquad i = 0,1,\cdots,N \tag{23}$$

设 $j \in R_1, W_j$ 表示迭加过程从开始处于相位 j 时起一直到它首次到达相位 0 时止这段时间，记 $\omega_j = E(W_j)$，则由全期望公式，有

$$\begin{aligned} \omega_j = E(W_j) &= E(Z_j) + E(W_j | X_j < Y_{N-j}]P\{X_j < Y_{N-j}\} \\ &\quad + E[W_j | X_j > Y_{N-j}]P\{X_j > Y_{N-j}\} \\ &= z_j + \lambda_j\omega_{j+1} + \mu_j\omega_{j-1} \qquad j = 1,2,\cdots,N-1 \\ \omega_N = E(W_N) &= E(Z_N) + \mu_N\omega_{N-1} = z_N + \omega_{N-1} \end{aligned}$$

即
$$\begin{cases} \omega_j = z_j + \lambda_j\omega_{j+1} + \mu_j\omega_{j-1}, j = 1,2,\cdots,N-1 \\ \omega_N = z_N + \omega_{N-1} \end{cases} \tag{24}$$

解此二阶差分方程，注意到边界条件 $\omega_0 = 0$，得

$$\omega_{j+1} - \omega_j = -\frac{z_j}{\lambda_j} + \frac{\mu_j}{\lambda_j}(\omega_j - \omega_{j-1}) \qquad \text{[递推]}$$

$$= -\frac{\mu_j\mu_{j-1}\cdots\mu_1}{\lambda_j\lambda_{j-1}\cdots\lambda_1}\omega_1 - \sum_{i=1}^{j}\frac{z_i}{\lambda_j}\prod_{k=i+1}^{j}\frac{\mu_k}{\lambda_k} \tag{25}$$

故
$$z_N = \omega_N - \omega_{N-1}) = \frac{\mu_{N-1}\mu_{N-2}\cdots\mu_1}{\lambda_{N-1}\lambda_{N-2}\cdots\lambda_1}\omega_1 - \sum_{i=1}^{N-1}\frac{z_i}{\lambda_i}\prod_{k=i+1}^{N-1}\frac{\mu_k}{\lambda_k}$$

从而
$$\omega_1 = \frac{\lambda_{N-1}\lambda_{N-2}\cdots\lambda_1}{\mu_{N-1}\mu_{N-2}\cdots\mu_1}z_N + \sum_{i=1}^{N-1}\frac{z_i}{\lambda_i}\cdot\frac{\lambda_i\lambda_{i-1}\cdots\lambda_1}{\mu_i\mu_{i-1}\cdots\mu_1}$$
$$= \rho_{N-1}z_N + \sum_{i=1}^{N-1}\frac{\rho_i z_i}{\lambda_i}$$

其中
$$\rho_i = \frac{\lambda_i\lambda_{i-1}\cdots\lambda_1}{\mu_i\mu_{i-1}\cdots\mu_1}, \qquad i = 1,2,\cdots,N-1 \tag{26}$$

将(26)式代入(25)式，得
$$\omega_{j+1} = \frac{1}{\rho_j}\omega_1 - \sum_{i=1}^{j}\frac{z_i}{\lambda_i}\prod_{k=i+1}^{j}\frac{\mu_k}{\lambda_k} + \omega_j \qquad [\text{递推}]$$
$$= \sum_{i=1}^{j}\frac{1}{\rho_i}\omega_1 + \omega_1 - \sum_{l=1}^{j}\sum_{i=1}^{l}\frac{z_i}{\lambda_i}\prod_{k=i+1}^{l}\frac{\mu_k}{\lambda_k}$$

故
$$\omega_j = \sum_{i=1}^{j-1}\frac{1}{\rho_i}\omega_1 + \omega_1 - \sum_{l=1}^{j-1}\sum_{i=1}^{l}\frac{z_i}{\lambda_i}\prod_{k=i+1}^{l}\frac{\mu_k}{\lambda_k} \tag{27}$$

其中规定：当 $i > j$ 时，$\sum_{k=i}^{j}a_k = 0$，$\prod_{k=i}^{j}a_k = 1$.

由(26)式知，当迭加过程由状态0转移到状态1时，它每次在状态1平均停留时间为 ω_1，而它在状态0每次平均停时为 $z_0 = \frac{1}{N\beta}$. 令
$$\frac{1}{\lambda} = z_0 + \omega_1, \quad \text{即} \quad \lambda = \frac{1}{z_0 + \omega_1} \tag{28}$$

我们把 $\frac{1}{\lambda} = z_0 + \omega_1$ 当成近似的(批到达)泊松过程中平均到达间隔时间。

由(26)式还可以看到，迭加过程每次在状态1平均停留时间为 ω_1，而 z_i 为迭加过程每次在相位 i 停留的平均时间。由(26)式易见，每次在状态1期间，它在相位 i 停留的总的平均时间为 $\frac{\rho_i z_i}{\lambda_i}$，$i = 1,2,\cdots,N-1$，在相位 N 停留的总平均时间为 $\rho_{N-1}z_N$，而 z_i 的系数 $\frac{\rho_i}{\lambda_i}$（z_N 的系数 ρ_{N-1}）即为迭加过程每次处于状态1期间经过相位 $i(N)$ 的平均次数。因此，每经过平均时间 $\frac{1}{\lambda} = z_0 + \omega_1$，平均到达的信息包数为
$$m = \sum_{i=1}^{N-1} i\cdot\frac{\rho_i z_i}{\lambda_i T} + N\cdot\frac{\rho_{N-1}z_N}{T} \tag{29}$$

令
$$r = [m] \tag{30}$$

这样近似的批到达泊松过程中的两个参数都已确定。

直观上易知，当 z_0 相对 ω_1 来说较大，即当 z_0 较大，ω_1 较小时近似程度应较好。或当 $\frac{1}{\alpha}$ 较小，$\frac{1}{\beta}$ 较大，且 N 不太大时近似程度应较好。

3　排队系统 $M^r/G/1$ 的解

注意到在排队系统 $M^X/G/1$ 中，$X \equiv r, x = r, \sigma_x^2 = 0, X(Z) = Z^r$ 和定理1,2,3，可得
$$P(Z) = \frac{(1-\rho)(1-Z)\tilde{V}(\lambda - \lambda Z^r)}{\tilde{V}(\lambda - \lambda Z^r) - Z}, \qquad \rho \equiv \frac{\lambda r}{\mu} < 1 \tag{31}$$
$$P'(1) = \rho + \frac{\rho(r^2 - r)}{2r(1-\rho)} + \frac{\rho^2(1+\mu^2\sigma^2)}{2(1-\rho)}, \qquad \rho < 1 \tag{32}$$

$$\tilde{W}_f(s) = \frac{s(1-\rho)}{s+\lambda-x[\tilde{V}(s)]^r}, \qquad \rho<1 \tag{33}$$

$$E(W_f) = \frac{\rho(r^2+r\mu^2\sigma^2)}{2r\mu(1-\rho)}, \qquad \rho<1 \tag{34}$$

$$\tilde{B}(s) = \{\tilde{V}[s+\lambda-\lambda\tilde{B}(s)]\}^r, \qquad \rho<1 \tag{35}$$

$$N(Z) = \{Z\tilde{V}[\lambda-\lambda N(Z)]\}^r, \qquad \rho<1 \tag{36}$$

$$E(B) = \frac{r}{\mu-\lambda r}, \quad D(B) = \frac{r^2\rho+r\mu^2\sigma^2}{\mu^2(1-\rho)^3}, \qquad \rho<1 \tag{37}$$

$$E(N) = \frac{r}{1-\rho}, \quad D(N) = \frac{r^2\rho+\lambda^2r^3\sigma^2}{(1-\rho)^3}, \qquad \rho<1 \tag{38}$$

式(31)～(38)中的 λ 由(28)式给出。

4 损失率

在解排队系统 $M^r/G/1$ 中，基于系统中的缓冲器是无限的的假设，没有考虑信息包的损失。但是实际中的缓冲器均是有限的，因此一般要考虑信息包的损失。设系统中缓冲器的最大容量为 L 个信息包(包括正在服务的信息包)。这样在第3节中不是求排队系统 $M^r/G/1$ 的解，而是要求排队系统 $M^r/G/1/L$ 的解。而所要求的损失率就是系统 $M^r/G/1/L$ 中有 L 个信息包的概率。更一般地，我们现求系统 $M^x/G/1/L$ 的解，其中 X 如第1节所设。

对于系统 $M^x/G/1/L$，当 $\rho \triangleq \frac{\lambda x}{\mu} < 1$ 时，由[6]知系统中有 k 个信息包的概率 p_k 由

$$p_k = p_0a_k + \sum_{j=1}^{k+1} p_ja_{k-j+1}, \quad 0 \leqslant k \leqslant L-2 \tag{39}$$

$$\sum_{k=0}^{L} p_k = 1 \tag{40}$$

$$p_0 = 1-\rho \tag{41}$$

给出，其中 a_k 为在一个顾客服务时间内到达 k 个信息包的概率，$k=0,1,2,\cdots$. 分布列 $\{a_k, k>0\}$ 的 p. g. f 为

$$\sum_{k=0}^{\infty} a_kz^k = \tilde{\Gamma}[(\lambda-\lambda X(Z)] \tag{42}$$

所以，损失率为 p_L. p_L 由(39)～(42)式确定。

参 考 文 献

1 H. Heffes and D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance", IEEE Journal on Selected Areas in Communications, Vol. SAC-4, NO. 6, September. 1986. 856～868

2 Blondie C. The N/G/1 Finite Capacity Queue, Commun. Stat-Stochast. Models, 1989, 5(2): 273～294

3 孟玉柯. 排队论基础及应用. 上海：同济大学出版社，1989. 232～240

4 Chaudhry M L, Templeton J G C. A First Course in Bulk Queues. New York: John Wiley & Sons Inc. 1983. 112～115

5 Yegenoglu F, Jabbari B. Performance Evaluation of MMPP/D/1/K Queues for Aggregate ATM Traffic Models. Proceedings of INFOCOM, 1993

6 Hideaki Takage. Queueing Analysis, Vol. 2, NORTH-HOLLAND AMSTERDAM. TOKYO; 1991. 412～426

1996 年 11 月 重庆大学学报 (自然科学版) Vol. 19, №. 6
第 19 卷第 6 期 Journal of Chongqing University (Natural Science Edition) Nov. 1996

(5) 27-32

前馈网误差反传的显式及网络的鲁棒性分析*

Explicit Formulation of the Back-Propagation and the Robustness Analysis of Feed-Forward Neural Networks

TP18

傅 鹂
Fu Li
(重庆大学系统工程及应用数学系,重庆,630044;35 岁,男,副教授)

摘 要 给出了前馈神经网络误差反传的一种紧凑显式表达,并据此对前馈神经网络的鲁棒性进行了分析,提出了设计鲁棒的前馈网所应遵循的若干参考原则。

关键词 神经网络;误差反传;鲁棒性
中国图书资料分类法分类号 TP302.7

前馈神经网络

ABSTRACT The compact explicit formulation of the Back-poopagation is presented, based on which the robustness of the feed-forward neural networks is analysed, and some reference priciples for designing a robust feed-forward net are provided.

KEYWORDS neural networks; back-propagation of error; robustness

0 引 言

在高度并行、模糊的大规模信息处理任务方面,(人工)神经网络有其独特的优点[1],因而在联想记忆、模式识别、自适应(自学习,自组织)控制等诸多领域中得到了广泛应用。尽管被研究提出的神经网络模型早已多达几十种[2],但前馈神经网络(Feed-Forward Neural Networds,以下简称"前馈网")仍然是最重要的网络之一。前馈网是早期感知器[3]的推广,由于其增加了隐单元层,使之具有普遍的函数逼近能力,从而能够完成某些十分复杂的信息处理任务。前馈网能力的实现是通过学习或训练,其中误差反传(BP:Back-Propagation)学习算法(BP 算法)占有极其重要的地位。

BP 算法的关键是计算误差梯度的反传表达式,而这个表达式的现有形式是十分松散的,它隐含在若干分量式中[4]。笔者导出了紧凑的显表达式。这一结果不仅使前馈网中误差反传机制以及 BP 算法的表达十分清晰,同时还为前馈网的鲁棒性分析提供了一种简明有效的途径,在这方面笔者亦给出了相应的研究结果,提出了设计鲁棒的前馈网应当遵循的若干原则。

1 误差反传的显式表达

* 收文日期 1996-03-15

1.1 前馈网的描述

考虑用一前馈网实现某个特定的映射关系 $\underline{v}=G(\underline{u})$，其中 $\underline{u}\in R^N$ 为输入向量，$\underline{v}\in R^M$ 为输出向量。网络参数包括联结权矩阵和阈值向量，为了能把阈值向量合并到联结权矩阵中，一般可在网络除输出层外的每一层中增添一个状态恒为 -1 的节点。这里采取在输出层中也同样增加这样一个状态恒为 -1 的节点，目的是使得有关的描述和处理更加统一，而不会改变问题的实质。若设 $\underline{x}=\begin{bmatrix}-1\\ \underline{u}\end{bmatrix}\in R^{N+1}$，$y=\begin{bmatrix}-1\\ \underline{v}\end{bmatrix}\in R^{M+1}$ 分别为增广的输入和输出向量，则网络实现的映射 $\begin{bmatrix}-1\\ \underline{v}\end{bmatrix}=\begin{bmatrix}-1\\ G(\underline{u})\end{bmatrix}$ 可表示为 $y=F(x)$，其中 $F:\begin{bmatrix}-1\\ \underline{u}\end{bmatrix}\to\begin{bmatrix}-1\\ \underline{v}\end{bmatrix}=\begin{bmatrix}-1\\ G(\underline{u})\end{bmatrix}$。下面给出这样一个前馈网内部结构的详细描述。

1) 网络共有 $L+1$ 层，分别标记为第 0 层，第 1 层，…，第 L 层（$L\geqslant 1$）；

2) 第 k 层共有 N_k+1 个节点，分别标记为节点 0，节点 1，…，节点 N_k （$N_k\geqslant 1;0\leqslant k\leqslant L$）；

3) 第 k 层的状态向量为

$$y_k=[-1,y_{k1},y_{k2},y_{kN_k}]^T\in R^{N_k+1}\qquad(0\leqslant k\leqslant L)\tag{1}$$

4) 第 0 层为输入层，第 L 层为输出层，故有

$$y_0\equiv x\in R^{N+1},\quad N_0\equiv N\tag{2}$$

$$y_L\equiv y\in R^{M+1},\quad N_L\equiv M\tag{3}$$

（第 $1,2,\cdots,L-1$ 为隐层（$L\geqslant 2$），若 $L=1$ 则没有隐层）；

5) 第 $k-1$ 层到第 k 层的状态传递规律为：

$$\begin{cases}x_k=W_k y_{k-1},\quad x_k=[x_{k0},x_{k1},\cdots,x_{kN_k}]^T\in R^{N_k+1},\quad W_k\in R^{(N_k+1)\times(N_{k-1}+1)}\\ y_k=f(x_k)\triangleq[-1,S(x_{k0}),S(x_{k1}),\cdots,S(x_{kN_k})]^T\in R^{N_k+1},\quad(1\leqslant k\leqslant L)\end{cases}\tag{4}$$

其中 x_k 为第 k 层的净输入向量，W_k 为第 $k-1$ 层到第 k 层的联结权矩阵（W_k 的第 0 列实际上代表第 k 层的阈值向量），$S(\cdot)$ 为节点激发函数，可采用 Sigmoid 函数如：

$$S(t)=\frac{1}{1+e^{-\alpha}}\qquad(\alpha>0)\tag{5}$$

1.2 BP 的显式表达

前馈网的输出误差函数为

$$E=\frac{1}{2}\|y-d\|^2\tag{6}$$

其中 $y\in R^{M+1}$ 是实际输出，$d\in R^{M+1}$ 是相应的期望输出，两者均由输入 $x\in R^{N+1}$ 决定；另一方面，E 也是联结权矩阵的函数，所以

$$E=E(x;W_1,W_2,\cdots,W_L)\tag{7}$$

学习算法的目的就是在某个输入样本集 $\{x_p|p\in P\}$ 上不断调整 $\{W_k|k=1,2,\cdots,L\}$ 使 E 的上界或均值达到最小。

BP 算法的实质是梯度法，其核心是通过误差反传求出 E 对联结权的导数。下面利用向量微分链规则[5] 直接导出紧凑的显式的 BP 公式。

记

$$D_k = \frac{\mathrm{d}y_k^T}{\mathrm{d}x_k} \qquad (0 \leqslant k \leqslant L) \tag{8}$$

由(2) 显然有

$$D_0 \equiv I \tag{9}$$

由(4) 不难得到 $D_k(1 \leqslant k \leqslant L)$ 为对角阵：

$$D_k = \mathrm{diag}[0, S'(x_{k0}), S'(x_{k1}), \cdots, S'(x_{kN_k})] \qquad (1 \leqslant k \leqslant L) \tag{10}$$

下面求 E 对联结权阵 $W_k(1 \leqslant k \leqslant L)$ 的导数矩阵：

$$\frac{\partial E}{\partial W_k} = \left[\frac{\partial E}{W_{ij}}\right]_{(N_k+1)\times(N_{k-1}+1)} \tag{11}$$

由(4) 知 $x_{ki} = \sum_j W_{kij} y_{k-1\ j}$，故

$$\frac{\partial x_{ki}}{\partial W_{kij}} = y_{k-1\ j} \tag{12}$$

于是有

$$\frac{\partial E}{\partial W_{kij}} = \frac{\partial E}{\partial x_{ki}} \frac{\partial x_{ki}}{\partial W_{kij}} = \frac{\partial E}{\partial x_{ki}} y_{k-1\ j}, \tag{13}$$

由

$$(ab^T)_{ij} = a_i b_j \tag{14}$$

(其中 a, b 为向量) 知(13) 的矩阵形式为

$$\frac{\partial E}{\partial W_k} = \frac{\partial E}{\partial x_k} y_{k-1}^T \tag{15}$$

由向量微分链规则[5] 有

$$\frac{\partial E}{\partial x_k} = \frac{\partial y^T}{\partial x_k} \frac{\mathrm{d}E}{\mathrm{d}y} = \frac{\partial y^T}{\partial x_k}(y - d) \tag{16}$$

其中用到了

$$\frac{\mathrm{d}E}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y}\left[\frac{1}{2} \| y - d \|^2\right] = y - d \tag{17}$$

再用链规则又有

$$\frac{\partial y^T}{\partial x_k} = \frac{\mathrm{d}y_k^T}{\mathrm{d}x_k} \frac{\partial x_{k+1}^T}{\partial y_k} \frac{\partial y^T}{\partial x_{k+1}} \tag{18}$$

由(8) 知 $\frac{\mathrm{d}y_k^T}{\mathrm{d}x_k} = D_k$，而由(4) 易得

$$\frac{\partial x_{k+1}^T}{\partial y_k} = W_{k+1}^T \tag{19}$$

故(18) 变为

$$\frac{\partial y^T}{\partial x_k} = D_k W_{k+1}^T \frac{\partial y^T}{\partial x_{k+1}} \tag{20}$$

将上式代回(16) 中得

$$\frac{\partial E}{\partial x_k} = D_k W_{k+1}^T \frac{\partial y^T}{\partial x_{k+1}}(y - d) \tag{21}$$

在(16)中将 k 换成 $k+1$ 可知 $\frac{\partial y^T}{\partial x_{k+1}}(y-d)=\frac{\partial E}{\partial x_{k+1}}$，于是(21)成为

$$\frac{\partial E}{\partial x_k}=D_kW_{k+1}^T\frac{\partial E}{\partial x_{k+1}} \qquad (1\leqslant k\leqslant L-1) \tag{22}$$

因 $\frac{\partial y^T}{\partial x_L}=\frac{\partial y_L^T}{\partial x_L}=D_L$，故由(16)

$$\frac{\partial E}{\partial x_L}=D_L(y-d) \tag{23}$$

将(22)代回(15)可得

$$\frac{\partial E}{\partial W_k}=D_kW_{k+1}^T\frac{\partial E}{\partial x_{k+1}}y_{k-1}^T \tag{24}$$

由(1)知 $\|y_k\|\geqslant 1(0\leqslant k\leqslant L)$，故

$$y_{k-1}^T\equiv\frac{\|y_k\|^2}{\|y_k\|^2}y_{k-1}^T=\frac{y_k^Ty_k}{\|y_k\|^2}y_{k-1}^T=y_k^T\left[\frac{y_ky_{k-1}^T}{\|y_k\|^2}\right] \tag{25}$$

由(15)得 $\frac{\partial E}{\partial x_{k+1}}y_k^T=\frac{\partial E}{\partial W_{k+1}}$，故将(25)代入(24)得

$$\frac{\partial E}{\partial W_k}=D_kW_{k+1}^T\frac{\partial E}{\partial W_{k+1}}\frac{y_ky_{k-1}^T}{\|y_k\|^2} \qquad (1\leqslant k\leqslant L-1) \tag{26}$$

而由(23)及(15)不难得出

$$\frac{\partial E}{\partial W_L}=D_L(y-d)y_{L-1}^T \tag{27}$$

到此已经得到所需的全部有关结果，下面总结为两种等价的 BP 公式：

第一种形式的 BP 公式(据(15)、(22)和(23))：

$$\begin{cases} z_L=D_L(y-d) \\ z_k=D_kW_{k+1}^Tz_{k+1}(k=L-1,L-2,\cdots,1) \\ \frac{\partial E}{\partial W_k}=z_ky_{k-1}^T \end{cases} \tag{28}$$

其中 $z_k=\frac{\partial E}{\partial x_k}$ 是第 k 层的误差信号，它在(28)中是一个中间变量，因为学习算法所直接需要的是输出误差函数对联结权的导数，即 $\frac{\partial E}{\partial W_k}$.

第二种形式的 BP 公式(直接由(26)、(27))：

$$\begin{cases} \frac{\partial E}{\partial W_L}=D_L(y-d)y_{L-1}^T \\ \frac{\partial E}{\partial W_k}=D_kW_{k+1}^T\frac{\partial E}{\partial W_{k+1}}\cdot\frac{y_k}{\|y_k\|^2}y_{k-1}^T \quad (k=L-1,L-2,\cdots,1) \end{cases} \tag{29}$$

上述第一种形式的BP公式(28)有利于计算，而第二种形式的BP公式(29)则更直接地表达了 $\frac{\partial E}{\partial W_k}$ 与 $\frac{\partial E}{\partial W_{k+1}}$ 的递推关系以及误差反传机制。

2 前馈网的鲁棒性分析

2.1　网络的鲁棒性与参数灵敏度矩阵

这里鲁棒性是指系统参数或结构变化时系统是否在一定误差范围内仍能正确输出。一个经过训练的前馈网，假设不考虑其结构的损坏(即拓扑结构不变)，而仅仅考虑参数的飘移(例如由元件老化引起)对网络的影响，则该网络的鲁棒性仅由输出对参数变化的灵敏度衡量，而这里参数就是联结权矩阵$\{W_k|k=1,2,\cdots,L\}$. 对参数的扰动可导致输出的变化，而这种变化对网络性能的影响应该由 $\|y-d\|$ 的变化大小来度量。因 $E=\frac{1}{2}\|y-d\|^2$，于是可以将$\frac{\partial E}{\partial W_k}$作为网络的参数灵敏度矩阵并用$\left\|\frac{\partial E}{\partial W_k}\right\|$来度量网络的鲁棒性。

由上节给出的两个 BP 公式中的任何一个均可得出上述参数灵敏度矩阵的明确表达式：

$$\begin{cases}\dfrac{\partial E}{\partial W_k}=D_k[D_LW_LD_{L-1}W_{L-1}\cdots D_{k+1}W_{k+1}]^T(y-d)y_{k-1}^T & (1\leqslant k\leqslant L-1)\\ \dfrac{\partial E}{\partial W_L}=D_L(y-d)y_{L-1}^T\end{cases}\tag{30}$$

这一式子为网络的鲁棒分析提供了直接途径。

2.2　鲁棒性分析及有关结论

令

$$\alpha=\max_t|S'(t)|\tag{31}$$

由于 $D_k=\mathrm{diag}[0,S'(x_{k1}),S'(x_{k2}),\cdots,S'(x_{kN_k})]$，故

$$\|D_k\|_p=\max_n|S'(x_{kn})|\leqslant\alpha\qquad(P=1,2,\infty)\tag{32}$$

由(30)可得

$$\begin{cases}\left\|\dfrac{\partial E}{\partial W_k}\right\|_p\leqslant\prod_{r=k}^{L}\|D_r\|_p\prod_{r=k+1}^{L}\|W_r\|_p\|y-d\|_p\|y_{k-1}\|_p, & (1\leqslant k\leqslant L-1)\\ \left\|\dfrac{\partial E}{\partial W_L}\right\|_p\leqslant\|D_L\|_p\|y-d\|_p\|y_{L-1}\|_p,\end{cases}\tag{33}$$

利用(32)上式成为

$$\begin{cases}\left\|\dfrac{\partial E}{\partial W_k}\right\|_p\leqslant\alpha^{L-k+1}\prod_{r=k+1}^{L}\|W_r\|_p\|y-d\|_p\|y_{k-1}\|_p, & (1\leqslant k\leqslant L-1)\\ \left\|\dfrac{\partial E}{\partial W_L}\right\|_p\leqslant\alpha\|y-d\|_p\|y_{L-1}\|_p,\end{cases}\tag{34}$$

为了将上式进一步具体化，假设采用(5)式的 Sigmoid 函数，并取 $p=\infty$ 范数(下面略去 ∞ 下标)，则由(4)知

$$\|y_k\|\leqslant 1\qquad(1\leqslant k\leqslant L)\tag{35}$$

于是(34)变为

$$\begin{cases} \left\|\dfrac{\partial E}{\partial W_k}\right\| \leqslant \alpha^{L-k+1} \prod_{r=k+1}^{L} \| W_r \| \, \| y - d \| & (1 \leqslant k \leqslant L-1) \\ \left\|\dfrac{\partial E}{\partial W_L}\right\| \leqslant \alpha \| y - d \| \end{cases} \tag{36}$$

根据(36)式可以得出鲁棒前馈网设计的有关参考原则：

1) 节点激发函数应较平缓，即各点斜率绝对值不应太大；

2) 联结权绝对值不应太大；

3) 网络训练应充分以使输出误差尽量小。

此外当 α 选得较大时，隐层数目应较小；而当 α 选得充分小时，则可增加隐层数目。

3 结 论

由(28)和(29)代表的两种 BP 公式，比常规形式更紧凑更清晰地描述了 BP 算法和前馈网中的误差反传机制，这有助于前馈网的分析研究，前述鲁棒性分析即为其中一个重要方面。

所提出的鲁棒性设计原则应当与其它原则一起综合应用，如网络容量、资源利用率等[6]. 再如，当激发函数斜率(绝对值)最大值 $\alpha \ll 1$ 时，根据前述鲁棒性分析，则网络隐层越多越好，这与常规似乎已经矛盾。其实，在实际中，α 不可能取得太小，否则网络的区分能力将下降，极限情况 $\alpha = 0$ 时，网络输出为恒值，根本没有区分能力。此外"联结权绝对值不应太大"这一点可通过在训练过程中加适当的限制而实现。一个简单有效的办法是在输出误差函数中增加 $\rho \sum \| W_i \|^2$ 项[7]，其中 $\rho > 0$ 反映对联结权绝对值限制的程度。

参 考 文 献

1 Soucek B, Soucek M. Neural and Massively Parallel Computers: the Sixth Generation. John Wiley & Sons, Inc. 1988. 48～69

2 Hecht-Nielsen R. Neurocomputing: Picking the human brain. IEEE Spectrum. IEEE, Inc. 1988, 25(3): 36～41

3 Rosenblatt R. Principles of Neuraldynamics. New York, Spartan Books, 1959. 55～70

4 Rumelbart D E, Hinton G E, Williams R J(Eds). Parallel Distributed Processing: Exploration in the Micro-structure of Cognition. Vol. 1: Foundations: MIT Press. 1986. 88～108

5 谢绪恺. 现代控制理论. 沈阳：辽宁人民出版社，1981. 221～229

6 张承福，赵刚. 联想记忆神经网络的若干问题. 自动化学报，1994，20(5)：513～521

7 Hinton G E. Connectionist Learning Procedures. Artificial Intelligence. 1989, 40: 185～234