第 19 卷 第 1 期
1993 年 1 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 19, No. 1
Jan., 1993

训练多层网络的样本数问题[1)]

张 鸿 宾
(北京计算机学院计算机科学系，100044)

摘 要

本文分析多层网络的映射增长函数，以经验风险最小和期望风险最小之间的偏差来定义网络的泛化能力。基于 Vapnik-Chervonenkis 的事件出现频率一致收敛于其概率的理论，讨论网络的结构、训练样本数和网络泛化能力间的关系。分析在最不利的情况下为保证一定泛化能力所需要的训练样本数。

关键词：人工神经网络，泛化，Vapnik-Chervonenkis 维数。

一、前 言

近年来，多层前馈网络的理论和应用研究取得了很大的进展。自从 Rumelhart 等人提出多层网络的 BP 学习算法以后，人们对多层网络的基本功能，它和函数逼近及多元分析间的关系，它和贝叶斯模式分类器、非参数概率密度估计以及和特征提取间的关系等问题都作了大量的研究工作，取得了一些新的研究成果。但是，在多层网络的研究中仍有许多基本问题尚未很好解决。例如，当中间层节点数有限时，多层网络是否存在最小二乘意义下的最优解，训练过的网络是否有泛化、一般化 (generalization) 能力等都是理论和实际应用上迫切需要解决的问题。

所谓网络的泛化问题是指，经过训练后的网络能否适应未经训练的工作样本，即网络装入学习样本的能力是否具有一般性的问题。从理论上看，多层网络的泛化能力不仅与学习样本的分布和工作样本的分布的偏差有关，而且和网络的结构以及学习样本数有关。最近，E. Baum 等人从计算学习理论的角度讨论了多层网络的泛化能力[1]。S. Miyake 等人从统计决策的角度分析了训练多层网络的样本数问题[2]。但是，Miyake 的论文对多层网络的映射函数的数量以及收敛速度的估计都有许多值得讨论和需要改进的地方。本文以 V. N. Vapnik 和 A. Ya. Chervonenkis 所提出的事件出现频率一致收敛于其概率的理论为基础[3,4]，从概率上定义网络的泛化能力，讨论网络的结构、学习样本数和泛化能力间的关系，分析在最不利情况下为保证一定泛化能力所需要的最少样本数。

二、Vapnik-Chervonenkis 维数和多层网络的映射增长函数

V. N. Vapnik 和 A. Ya. Chervonenkis 在研究模式识别中经验风险最小和期望

本文于 1991 年 11 月 11 日收到。
1)国家自然科学基金资助的课题。本文的部分内容曾在中国神经网络 1991 年学术大会上宣读。

风险最小之间的关系时，曾经引入了判别函数类的容量的概念．这个概念现在一般称为VC维数．VC维数是一个非常重要的概念，在模式识别和计算学习理论等领域中有着广泛的应用．结合到多层网络，有如下定义．

定义． 假定样本 (x,ω)，$x\in R^n$，$\omega\in\{0,1\}$ 是随机地从 $R^n\times\{0,1\}$ 上的某个概率分布中抽出的．F 是从 $R^n\to\{0,1\}$ 的某种函数类．令 s 表示 R^n 中 m 点的集合，$d_F(s)$ 表示由 $f\in F$ 对 s 所产生的不同的二分割 (dichotomy) 数．令 $d_F(m)=\max\limits_{s\in s'} d_F(s)$，其中 s' 是所有点数为 m 的 $s\subset R^n$ 的集合．当 $d_F(m)=2^m$ 时，s 称为被 F 所细分的 (shattered)．函数类 F 的VC维数 VC(F) 定义为能被 F 所细分的 s 的最大元素数，即能使 $d_F(m)=2^m$ 的最大的 m．当 $d_F(m)=2^m$ 对任意的 m 都成立时，这时 F 的VC维数称为无穷大．$d_F(m)$ 称为函数类 F 的增长函数 (growth function)．

在上述定义中，如果 F 是线性函数类，由于 R^n 中的点集和平面集合间的对偶性，$d_F(m)$ 也表示 m 个超平面把 n 维空间分割的最多块数．

为了讨论多层网络的泛化能力，首先分析多层网络的映射增长函数 $d_F(m)$，它对网络所需要的训练样本数有直接的影响．

引理 1. 假定三层前馈网络有 n 个输入节点，h 个隐节点，一个输出节点．隐节点和输出节点的输入输出函数 g 为线性阈值型，即某个节点 i 的输出 $o_i=g\left(\sum\limits_j w_{ij}x_{ij}-\theta_i\right)$，其中当 $z\geqslant 0$ 时，$g(z)=1$；当 $z<0$ 时，$g(z)=0$．整个网络形成一个映射函数 $f:R^n\to\{0,1\}$．设 F 是网络所能产生的函数类，则有 $d_F(m)\leqslant 2^k\cdot m^{hn}$，其中 m 是训练样本数，而 $k=\sum\limits_{i=0}^{n} c_h^i$．

证明．令 $\phi(m,n)$ 表示 m 个超平面分割 n 维空间时所形成的最多区域数，则有[5]

$$\phi(m,n)=\sum_{i=0}^{n} c_m^i.$$

容易验证，当 $m\geqslant 0$ 和 $n\geqslant 0$ 时，有 $\phi(m,n)\leqslant m^n+1$；当 $m\geqslant 2$ 且 $n\geqslant 2$ 时，有 $\phi(m,n)\leqslant m^n$．

对于 m 个训练样本，每个超平面所形成的二分割数小于或等于 m^n．h 个平面所形成的分割数小于或等于 m^{hn}．又由于 h 个超平面把 n 维空间最多分割为

$$\phi(h,n)=\sum_{i=0}^{n} c_h^i=k$$

个区域，每个区域可以对应于0或1．因此有

$$d_F(m)\leqslant 2^k\cdot m^{hn}. \tag{1}$$

三、多层网络学习的泛化能力

多层网络学习时，本质上是求出一组适当的权 $\boldsymbol{w}$，使网络形成的映射函数 $f(x,\boldsymbol{w})$ 能使下式的泛函达到最小：

$$P(\boldsymbol{w}) = \int (\omega - f(x,\boldsymbol{w}))^2 p(x,\omega)\,dx\,d\omega, \tag{2}$$

式中 $p(x,\omega)$ 是 x 和 ω 的联合概率密度．$(\omega - f(x,\boldsymbol{w}))^2$ 是损失函数，输出正确时损失为 0，错误时损失为 1．(2)式实际上是函数 $f(x,\boldsymbol{w})$ 的期望风险，表示 $f(x,\boldsymbol{w})$ 的错分概率．

由于一般不知道 $p(x,\omega)$，而只是知道一组样本

$$(x_1,\omega_1),(x_2,\omega_2),\cdots,(x_m,\omega_m), \tag{3}$$

可以利用这组样本先估计概率密度，然后利用估计出的 $\hat{p}(x,\omega)$ 求泛函(2)式的极值． 但一般情况下概率密度的估计是一个更困难的问题．因此实际上很少采用．通常的做法是利用(3)式的样本求下面泛函的极值：

$$V(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m}(\omega_i - f(x_i,\boldsymbol{w}))^2. \tag{4}$$

(4)式是函数 $f(x,\boldsymbol{w})$ 的经验风险，表示 $f(x,\boldsymbol{w})$ 的错分频率．

网络学习时，可以利用 BP 等算法，求出一组使 $V(\boldsymbol{w})$ 最小的 $\boldsymbol{w}$．这里的问题是，使(4)式最小的 $f(x,\boldsymbol{w}_{exp})$ 是否等于或接近使(2)式最小的 $f(x,\boldsymbol{w}_0)$．

根据经典概率理论，当样本数趋于无穷时，有

$$\lim_{m\to\infty} P\{|P(\boldsymbol{w}) - V(\boldsymbol{w})| > \varepsilon\} = 0. \tag{5}$$

但(5)式并不意味着使经验风险最小的 $f(x,\boldsymbol{w}_{exp})$ 也使或接近使期望风险最小．如图 1 所示，使 $V(\boldsymbol{w})$ 最小的权 $\boldsymbol{w}_{exp}$ 和使 $P(\boldsymbol{w})$ 最小的权 $\boldsymbol{w}_0$ 间可能有很大的偏差．

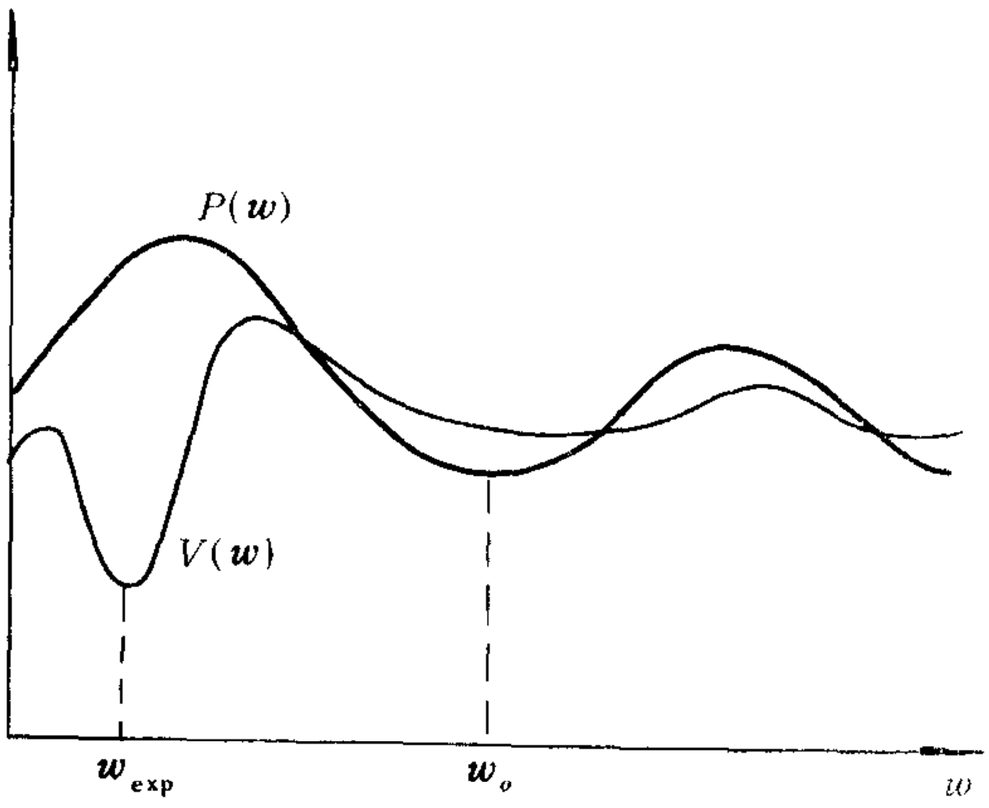

图 1　经验风险最小和期望风险最小之间的偏差

Vapnik 和 Chervonenkis 指出，为保证经验风险最小充分接近期望风险最小，应当要求更强的条件，即

$$\lim_{m\to\infty} P\{\sup_i |P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)| > \varepsilon\} = 0. \tag{6}$$

当(6)式满足时，称经验均值以参数 $\boldsymbol{w}$ 一致收敛于其数学期望．关于 Vapnik 和 Chervonenkis 的理论，文献[6]中有扼要的介绍．

在实际工作中，由于样本数有限，可以要求按一定概率保证经验风险最小充分接近期望风险最小．即要求

$$P\{\sup_{\boldsymbol{w}} |P(\boldsymbol{w}) - V(\boldsymbol{w})| > \varepsilon\} < \eta(m,\varepsilon),$$
$$\lim_{m\to\infty}\eta(m,\varepsilon) = 0. \tag{7}$$

(7)式等价于对所有的 $\boldsymbol{w}$，下面的区间以概率 $1-\eta(m,\varepsilon)$ 成立，

$$V(\boldsymbol{w}) - \varepsilon \leqslant P(\boldsymbol{w}) \leqslant V(\boldsymbol{w}) + \varepsilon. \tag{8}$$

由于 $\eta(m,\varepsilon)$ 是 m 和 ε 的减函数，因此对于给定的置信水平 $1-\eta$，其中

$$\eta = \eta(m,\varepsilon), \tag{9}$$

则由(8)式得到的置信区间 $\varepsilon = \varepsilon(m,\eta)$ 将随 m 的增大而减小。因此当 m 足够大时，使经验风险最小的 $\boldsymbol{w}_{exp}$ 将接近使期望风险最小。而对于固定的样本大小，权 $\boldsymbol{w}_{exp}$ 将以概率 $1-\eta$ 保证期望风险在下式的区间之内：

$$V(\boldsymbol{w}_{exp}) - \varepsilon \leqslant P(\boldsymbol{w}_{exp}) \leqslant V(\boldsymbol{w}_{exp}) + \varepsilon. \tag{10}$$

上述一致收敛的理论为分析多层网络学习的泛化问题提供了很好的工具。假定学习样本和工作样本抽自同一分布，那么网络的泛化能力可以用 $V(\boldsymbol{w})$ 以精度 ε 接近 $P(\boldsymbol{w})$ 的概率来衡量。它与置信水平 $1-\eta$、训练样本数以及网络的映射增长函数有关。

四、训练多层网络的样本数分析

下面将区别几种不同的情况，讨论训练多层网络的样本数问题。

1. 对所有事件类的一致收敛

假定多层网络的映射增长函数取有限值 N。网络的映射函数为

$$f(x,\boldsymbol{w}_1), f(x,\boldsymbol{w}_2), \cdots, f(x,\boldsymbol{w}_N).$$

对于每一个函数 $f(x,\boldsymbol{w}_i)$，可以定义一个事件 A_i，它由使 $(\omega - f(x,\boldsymbol{w}_i))^2 = 1$ 的样本 (x,ω)所组成。对每一个固定的事件，根据大数定律，下面的 Hoeffding 不等式成立，

$$P\{|P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)| > \varepsilon\} < 2e^{-2\varepsilon^2 m}. \tag{11}$$

由于

$$P\{\sup_i |P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)| > \varepsilon\} \leqslant \sum_{i=1}^{N} P\{|P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)| > \varepsilon\}, \tag{12}$$

因此有

$$P\{\sup_i |P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)| > \varepsilon\} \leqslant 2Ne^{-2\varepsilon^2 m}. \tag{13}$$

由(13)式，当 $m \to \infty$ 时，由于

$$\lim_{m\to\infty} P\{\sup_i |P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)| > \varepsilon\} = 0, \tag{14}$$

因此，对上述的有限事件类(有限的网络增长函数)，事件出现频率一致收敛于其概率。

对于有限的样本数，可以要求

$$P\{\sup_i |P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)| > \varepsilon\} < \eta. \tag{15}$$

由(13)式，令

$$\eta = 2Ne^{-2\varepsilon^2 m}, \tag{16}$$

从中解出 ε，可以得到事件类中频率对概率的偏差的估计：

$$\varepsilon = \left(\frac{\ln N - \ln(\eta/2)}{2m}\right)^{1/2}. \tag{17}$$

如果从(16)式中解出 m，那么可以得到样本的大小 m，它保证了频率对概率的偏差不超过 ε 的概率至少为 $1-\eta$，

$$m = \frac{\ln N - \ln(\eta/2)}{2\varepsilon^2}. \tag{18}$$

由以上分析，可得下面的定理.

定理 1. 如果网络的映射增长函数为有限值 N，在训练样本数为 m 时，网络函数 $f(x, w_i)$ 的错误频率记为 $V(w_i)$，则下面的不等式以概率 $1-\eta$ 对所有的网络函数都成立：

$$V(w_i) - \left(\frac{\ln N - \ln(\eta/2)}{2m}\right)^{1/2} \leqslant P(w_i) \leqslant V(w_i) + \left(\frac{\ln N - \ln(\eta/2)}{2m}\right)^{1/2}. \quad (19)$$

由于(19)式对所有的网络函数都成立，因此对使经验风险最小的网络函数 $f(x, w_{exp})$，可得下面的置信区间：

$$V(w_{exp}) - \left(\frac{\ln N - \ln(\eta/2)}{2m}\right)^{1/2} \leqslant P(w_{exp}) \leqslant V(w_{exp}) + \left(\frac{\ln N - \ln(\eta/2)}{2m}\right)^{1/2}. \quad (20)$$

2. 对特定事件类的一致收敛

定理 1 对置信区间的估计有可能过宽. 如果有关于网络函数的先验信息，那么可以改进上述的估计.

假定在网络的 N 个函数中，有一些函数的错分率不超过 $(1-r)\varepsilon$，$0 < r \leqslant 1, 0 < \varepsilon < 1$. 因此，通过任意 m 个训练样本有可能找到错分率最大为 $(1-r)\varepsilon$ 的网络函数. 但是，由于错分率超过 $(1-r)\varepsilon$ 的函数也可能对 m 个训练样本的错误率低于 $(1-r)\varepsilon$，因此有必要估计这样求得的函数其实际错误概率超过 ε 的概率.

引入函数

$$\theta(z) = \begin{cases} 1, & \text{当 } z \leqslant 0 \text{ 时}, \\ 0, & \text{当 } z > 0 \text{ 时}, \end{cases}$$

这时只要估计下式的概率：

$$\{\sup_i (P(w_i) - V(w_i)) \cdot \theta(V(w_i) - (1-r)\varepsilon) > r\varepsilon\}. \quad (21)$$

由于错分频率不大于 $(1-r)\varepsilon$ 的函数个数不会超过总数 N，因此有

$$P\{\sup_i (Pw_i) - V(w_i)) \cdot \theta(V(w_i) - (1-r)\varepsilon) > r\varepsilon\} \leqslant N \cdot P_\varepsilon^m. \quad (22)$$

式中 P_ε^m 是错分率等于 ε、但对 m 个样本的错分频率却不大于 $(1-r)\varepsilon$ 的概率. 显然，

$$P_\varepsilon^m = \sum_k \binom{m}{k} \cdot \varepsilon^k \cdot (1-\varepsilon)^{m-k}, \quad k \leqslant (1-r)\varepsilon m. \quad (23)$$

可以证明[7]，$P_\varepsilon^m \leqslant e^{-r^2 m\varepsilon/2}$. 因此，

$$P\{\sup_i (P(w_i) - V(w_i)) \cdot \theta(V(w_i) - (1-r)\varepsilon) > r\varepsilon\} \leqslant Ne^{-r^2\varepsilon m/2}. \quad (24)$$

令 $\eta = Ne^{-r^2\varepsilon m/2}$，可得

$$m = \frac{2(\ln N - \ln\eta)}{r^2\varepsilon}, \quad (25)$$

$$\varepsilon = \frac{2(\ln N - \ln\eta)}{r^2 m}. \quad (26)$$

由(25)式可以看出，这时的样本大小与 $1/\varepsilon$ 成正比，而不是(18)式的 $1/\varepsilon^2$. 因此需要的样本数减少了. 由此可得下面的定理.

定理 2. 如果从 N 个网络函数中选择了一个函数，它对 m 个样本的错分频率不超过

$(1-r)\varepsilon$，那么该函数的错分概率 P 在下述区间内的概率大于 $1-\eta$，即

$$0 \leqslant P \leqslant \varepsilon. \tag{27}$$

式中 ε 和 N, m，以及 η 间的关系满足(26)式.

3. 和最优映射函数的偏差

在实际工作中，常常需要估计用经验风险最小所得到的网络函数 $f(x, \boldsymbol{w}_1)$ 和网络的最优函数 $f(x, \boldsymbol{w}_0)$ 之间的偏差. 如图 2 所示，一般要求根据经验风险最小所得到的 $\boldsymbol{w}_1$ 所对应的 $P(\boldsymbol{w}_1)$ 在下述范围内就可以了，

$$P(\boldsymbol{w}_1) \leqslant P(\boldsymbol{w}_0) + \varepsilon, \tag{28}$$

ε 表示 $P(\boldsymbol{w}_1)$ 和 $P(\boldsymbol{w}_0)$ 的接近程度. 由于对固定的 $\boldsymbol{w}$，频率收敛于概率的速度比一致收敛的速度要快得多，只要学习样本达到一定大小时，就可以认为 $P(\boldsymbol{w}_0) \approx V(\boldsymbol{w}_0)$. 如果用 ε 表示 $f(x, \boldsymbol{w})$ 和 $f(x, \boldsymbol{w}_0)$ 的接近程度，那么只要对所有 $P(\boldsymbol{w}) > P(\boldsymbol{w}_0) + \varepsilon$ 的 $\boldsymbol{w}, V(\boldsymbol{w})$ 比 $V(\boldsymbol{w}_0)$ 大，就能保证网络学习的品质. 从上述分析出发，根据 Vapnik-Chervonenkis 的单侧相对偏差估计[4]：

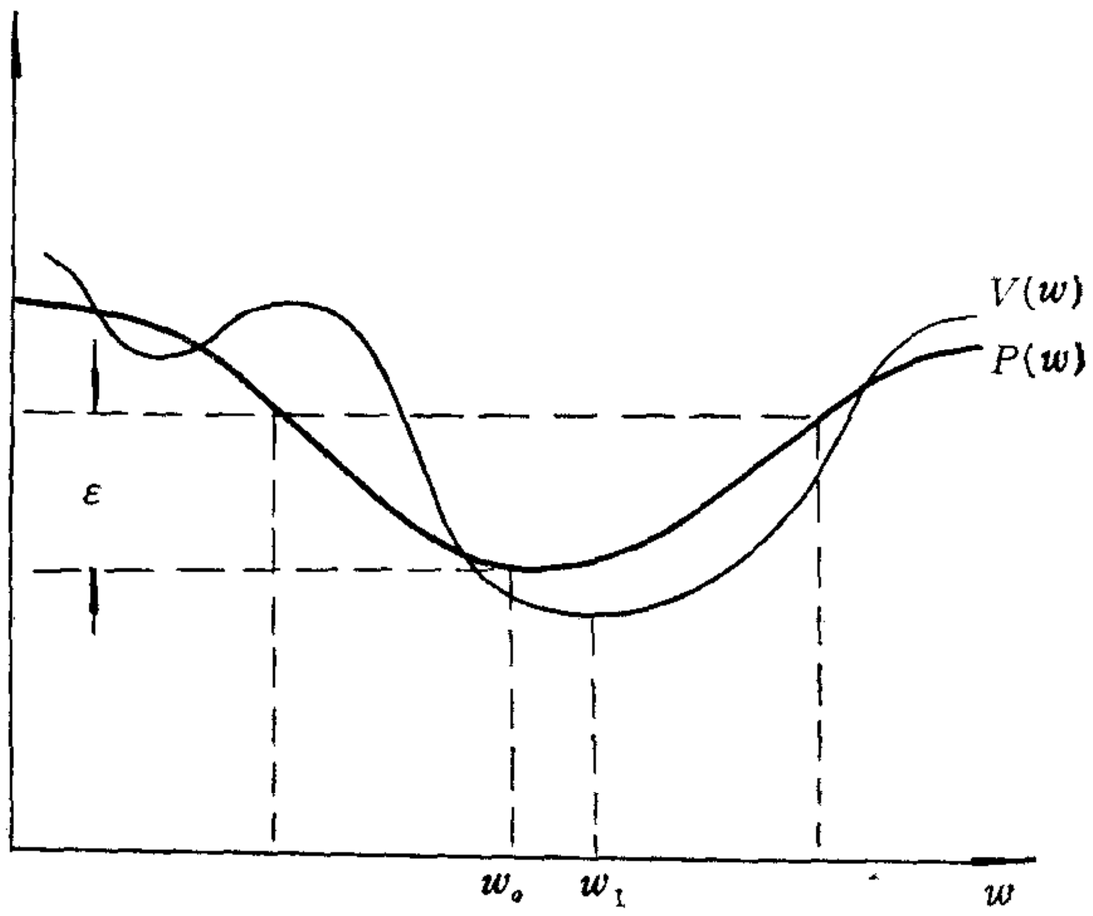

图 2 $P(\boldsymbol{w}_1)$ 所允许的误差范围

$$P\left\{\sup_i \frac{P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)}{\sqrt{P(\boldsymbol{w}_i)}} > \delta\right\} < 8 \cdot d_F(2m) e^{-\delta^2 m/4}. \tag{29}$$

令 $\delta = \dfrac{\varepsilon}{\sqrt{P(\boldsymbol{w}_0) + \varepsilon}}$，当满足条件

$$\sup_i \frac{P(\boldsymbol{w}_i) - V(\boldsymbol{w}_i)}{\sqrt{P(\boldsymbol{w}_i)}} \leqslant \delta \tag{30}$$

时，有

$$V(\boldsymbol{w}_i) \geqslant P(\boldsymbol{w}_i) - \frac{\varepsilon}{\sqrt{P(\boldsymbol{w}_0) + \varepsilon}} \cdot \sqrt{P(\boldsymbol{w}_i)}. \tag{31}$$

由于 $P(\boldsymbol{w}_i) > P(\boldsymbol{w}_0) + \varepsilon$，因此有 $V(\boldsymbol{w}_i) > P(\boldsymbol{w}_0) \approx V(\boldsymbol{w}_0)$. 由以上的分析可得：

$$P\{(P(\boldsymbol{w}_1) - P(\boldsymbol{w}_0)) > \varepsilon\} < 8 d_F(2m) \cdot e^{\frac{-\varepsilon^2 m}{4(P(\boldsymbol{w}_0)+\varepsilon)}}. \tag{32}$$

令 $d_F(2m)=N$，$8d_F(2m)\cdot e^{\frac{-\varepsilon^2 m}{4(P(w_0)+\varepsilon)}}=\eta$，解出 m 得：

$$m=\frac{4(\ln N-\ln(\eta/8))\cdot(P(w_0)+\varepsilon)}{\varepsilon^2}. \tag{33}$$

(33)式表示了 $P(w_0)$，$d_F(2m)$，η，ε 和 m 间的关系．如果网络的隐节点数足够多的话，多层网络分类器可以充分逼近贝叶斯分类器．这时的 $P(w_0)$ 可以表示贝叶斯分类器的错误率．由式(32)可得下面的定理．

定理 3. 当网络函数中含有贝叶斯最优判别函数时，用(33)式的有限样本 m 训练的网络，其错分概率以 ε 精度接近贝叶斯错误率的概率大于 $1-\eta$．

参　考　文　献

[1] Baum, E. and Haussler, D., What Size Net Gives Valid Generalization?, Neural Computation, **1**(1989), 1, 151—160.

[2] *Kanaya, F. and Miyake, S.*, Bayes Statistical Behavior and Valid Generalization of Pattern Classifying Neural Networks, *IEEE Trans. Neural Networks*, **2**(1991), 4, 471—475.

[3] Vapnik, V. N. and Chervonenkis, A. Ya., On the Uniform Convergence of Relative Frequence of Events to their Probabilities, Theoret. Prob. and Its Appl., **16**(1971), 2, 264—280.

[4] Vapnik, V. N. and Chervonenkis, A. Ya., Theory of Pattern Recognition (in Russian), Nauka, Moscow, 1974.

[5] Nilsson, N. J., Learning Machine, New York: McGraw-Hill, 1965.

[6] 边肇祺等，模式识别，清华大学出版社．

[7] Angluin, D. and Valiant, L. G., Fast Probabilistic Algorithms for Hamiltonian Circuits and Matching, *J. Comput. Syst. Sci.*, **18**(1979), 155—193.

THE SAMPLE SIZE FOR TRAINING MULTI-LAYERED NEURAL NETWORK

ZHANG HONGBIN

(*Dept. of Computer Science, Beijing Computer Institute, Beijing,* 100044)

ABSTRACT

Based on the theory of uniform convergence of frequencies of events to their probabilities given by Vapnik and Chervonenkis, the relation-ship among the number of mapping functions, the size of training samples, and the ability of generalization of the multilayered neural network is discussed. The minimum training sample size, which guarantees valid generalization in the worst case, is analysed.

Key words: Neural network; generaliization; Vapnik-Chervonenkis dimenstion.

张鸿宾　1968 年清华大学自动控制系毕业．1981 年清华大学模式识别与智能控制专业硕士研究生毕业．毕业后在北京计算机学院从事教学和科研工作．副研究员．1986 年—1989 年在日本京都大学做访问学者．目前感兴趣的研究领域有模式识别、图象处理与分析、计算机视觉和人工神经网络等．