BP 网络学习能力与泛化能力满足的不确定关系式*

李祚泳[1] 彭荔红[2**]

(1. 成都信息工程学院, 成都 610041; 2. 厦门大学环境科学研究中心, 厦门 361005)

摘要 分析 BP 网络过拟合出现时网络学习能力及泛化能力与其他影响因素之间的内在联系, 引入复相关系数描述样本复杂性程度; 遵循计算不确定性原理和神经网络结构设计的最简原则, 类比信息传递过程中的一般测不准关系式, 建立了 BP 网络过拟合出现时, 反映网络学习能力的训练样本集的训练相对误差与表征泛化能力的网络对检验样本集的测试相对误差之间满足的不确定关系式; 通过模拟多种不同类型函数的 BP 网络过拟合数值模拟实验, 确定了关系式中过拟合参数 q 的取值范围一般为 $7\times10^{-3}\sim 7\times10^{-2}$; 依据不确定关系式, 导出了在用复相关系数描述样本复杂性和满足给定逼近误差要求下, 网络具有较佳泛化能力的隐节点数的计算公式, 并验证了其合理性; 指出 BP 网络应用于给定样本集的训练过程中, 为改进泛化能力的训练最佳停止方法.

关键词 **BP 网络 学习能力 泛化能力 过拟合关系式 网络结构优化**

BP 网络存在过拟合现象. 所谓过拟合是指只要允许网络足够复杂, BP 网络可使训练样本集的误差减少到足够小, 但当样本数有限情况下, 却会导致网络的泛化能力下降. 可见, 学习能力与泛化能力是有联系的. 因此, 研究 BP 网络的学习能力与泛化能力之间的关系是解决过拟合现象的关键. 这个问题的解决将推动 BP 网络理论和应用的发展. 正因为如此, 网络的学习能力与泛化能力的相互关系及其他因素对它们的影响的研究愈来愈受到人们的关注. 关于网络的泛化能力的研究, 国内外已出现了不少工作[1~14]. 对于单隐层全连接前向网络, Baum 得出了为保证固定结构的神经网络泛化能力所需训练样本数的上界和样本复杂性的下界[1]. Moody 通过研究测试集泛化误差与训练集误差之间关系, 得出了实值神经网络结构设计的最简原则[2]. Barron 等给出了学习后神经网络的泛化误差同样本数和隐节点数之间的关系, 得出随着隐节点数增加, 逼近误差逐渐减小, 而估计误差逐步增大, 好的泛化能力取决于二者的协调[3,4]. Cataltepe 等把所有样本数据分为训练数据和测试数据, 并当测试误差达到最小时, 停止网络的训练, 用以改进神经网络的泛化能力[5]. Amari 等研究了交叉测试法中测试样本数占总样本数比例对神经网络泛化能力的影响[6]. Baldi 对线性网络的训练和泛化动态进行了研究, 发现初始权值对神经网络的泛化能力有很大影响[7]. Partridge 对三层 BP 网络研究发现训练样本集对网络泛化能力亦有较大影响[8]. 张鸿滨讨论了为保证多层前向网络的泛化能力所需的

2003-03-04 收稿, 2003-07-04 收修改稿

* 国家自然科学基金(批准号: 40271024)及"九五"国家重点科技攻关(96-911-08-03)资助项目

** E-mail: lhpeng@xmu.edu.cn

样本数问题[9]. 魏海坤和阎平凡等分析了多层前向网络的泛化能力与结构复杂性和样本复杂性的关系[10,11]. 王晖等提出了基于主要影响因素和基于修正误差函数来增强网络泛化能力的方法, 提高了 BP 网络的适应性[12]. 江学军等认为影响前馈网络泛化能力的根本原因是训练网络用的样本质量、样本数量和样本代表性三个方面, 指出了一种提高泛化能力的途径[13]. 彭汉川等从提高网络特征提取能力、分类能力和修改激活函数等方面给出了提高前馈网络泛化能力的若干实际方案[14]. 不过已有的研究大多集中于分析训练样本复杂性或网络结构的复杂性对网络泛化能力的影响, 而对 BP 网络的学习能力及泛化能力与网络结构、样本复杂性(包括样本规模: 样本数 m 和样本因子数 n; 模拟函数复杂性)之间关系的定量研究则未见报道.

本文将 BP 网络学习及测试过程与信息传递过程进行类比分析, 由信息传递过程中的一般测不准关系式, 建立了 BP 网络过拟合时, 表示学习能力的训练样本集的训练相对误差及表示泛化能力的检验样本集的测试相对误差与网络结构的隐节点数和表征样本复杂性的复相关系数之间满足的一般不确定关系式; 并根据此关系式, 导出对给定样本集, 在满足逼近误差要求下, 遵循网络结构设计最简原则的网络学习能力与泛化能力相协调的最佳隐节点数计算公式; 最后指出为使网络有好的泛化能力的训练最佳停止法.

1 学习能力及泛化能力与其他因素之间满足的不确定关系式

在信息论中, 系统传递的最大平均信息量 S 满足一般测不准关系式[15]

$$|\Delta p|\cdot|\Delta x| \geqslant \frac{S\hbar}{2\log_2(1+M/N)}, \tag{1}$$

式中: Δp 和 Δx 分别为物体的动量变化和位置变化; S 为传递的最大平均信息量; M 为信息的平均功率; N 为环境噪声的平均功率; M/N 为信噪比, $\hbar = h/2\pi$, h 为普朗克常数.

从信息论观点看, BP 网络的学习和测试过程可视作信息的传递过程. 一定结构的 BP 网络就是一个信息传递系统. 所谓 BP 网络的学习和测试过程就是针对不同的网络结构, 不断调整网络的权值、阈值等参数, 去逼近模拟函数, 使逼近误差尽可能地小. 但事实是: 无论设计何种网络结构和怎样调整参数, 当学习进行到一定阶段后, 若学习样本集的平均训练相对误差继续减小, 而检验样本集的平均测试相对误差(泛化误差)反而增大, 即出现了过拟合现象. 也就是说 BP 网络过拟合时, 其学习能力与泛化能力之间的变化是反向的, 只有在过拟合刚开始点, 它们之间才有最好的匹配. 因此, 过拟合出现时, 学习样本集的平均训练相对误差与检验样本集的平均测试相对误差之间满足一个类似(1)式的不确定关系式. 设 M 和 N 分别为训练样本集的网络输出平均归一化值和输出均方根误差, 则它们的比值 M/N 代表了网络输出的信噪比. 此外, BP 网络传递的信息量大小应与网络结构的隐节点数 H, 问题规模(训练样本数 m 和样本因子数, 即网络输入节点数 n)和问题复杂性(自变量和因变量之间的函数类型)都有关, 而问题规模和问题复杂性可用复相关系数 R 近似描述. 复相关系数 R 描述自变量因子集综合起来与因变量关系的密切程度, 是反映自变量因子集合优劣程度的数量指标. n 个因子与因变量 y 的复相关系数 R 可用单相关系数求得[16]:

$$R = \sqrt{1 - R^* / R^*_{yy}}, \tag{2}$$

式中, R^* 和 R^*_{yy} 分别是单相关系数矩阵

$$R_{(n+1)\times(n+1)}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} & r_{1y} \\ r_{21} & r_{22} & \cdots & r_{2n} & r_{2y} \\ \vdots & \vdots & & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} & r_{ny} \\ r_{y1} & r_{y2} & \cdots & r_{yn} & r_{yy} \end{bmatrix} \tag{3}$$

的行列式及其去掉第 $n+1$ 行, 第 $n+1$ 列后的代数余子式值.

规定 R 取正值, 故有 $0\leqslant R\leqslant 1$, 多元线性函数的 $R=1$, 愈接近线性的函数的 R 愈接近 1, 愈非线性的复杂函数的 R 愈接近 0. 研究还表明: 复相关系数 R 随因子数 n 增加而增大, 随样本数 m 的增大而减小. 因此复相关系数 R 不仅能反映函数或样本集合本身性质的复杂性, 同时也包含了问题规模(样本数量 m 和因子数 n)的复杂性.

达到给定泛化能力的 BP 网络传递的信息量 S 应遵循神经网络结构设计的最简原则: 必须使网络结构的复杂性(主要指神经网络的隐节点数 H)与问题规模(主要指训练样本数 m)相匹配. 其选择只能是: 要么增大训练样本数 m(即减小复相关系数 R), 要么减小 BP 网络结构(即减少网络的隐节点数 H)[10]. 故满足网络结构的复杂性与问题规模相匹配的最简设计原则的 BP 网络所传递的信息量 S 可设定为 $S=RH$. 因此, 与信息传递过程中的测不准关系式(1)相类似, BP 网络过拟合出现时, 网络对训练样本集的平均训练相对误差 $\left|\overline{\dfrac{\Delta z}{z}}\right|$ 及网络对检验样本集的平均测试相对误差 $\left|\overline{\dfrac{\Delta y}{y}}\right|$ 与上述有关量之间也应满足如下的过拟合不确定关系式:

$$\left|\overline{\frac{\Delta z}{z}}\right|\cdot\left|\overline{\frac{\Delta y}{y}}\right|\geqslant\frac{RHq}{2\log_2(1+M/N)}, \tag{4}$$

式中, R 是描述问题复杂性和问题规模(样本集复杂性)的复相关系数, H 是网络的隐节点数, M/N 是训练样本集的网络平均输出值与方均根误差之比, q 为待定的“过拟合参数”. $\left|\overline{\dfrac{\Delta z}{z}}\right|$, $\left|\overline{\dfrac{\Delta y}{y}}\right|$, M 和 N 分别用以下式子计算:

$$\left|\overline{\frac{\Delta z}{z}}\right|=\frac{1}{m}\sum_i\left|\frac{O_i-T_i}{T_i}\right|, \tag{5}$$

$$\left|\overline{\frac{\Delta y}{y}}\right|=\frac{1}{m'}\sum_{i'}\left|\frac{O_{i'}-T_{i'}}{T_{i'}}\right|, \tag{6}$$

$$M=\sum O_{ki}/m, \tag{7}$$

$$N=\sqrt{\sum_i^m(O_{ki}-T_{ki})^2/m}, \tag{8}$$

式中, m 和 m' 分别为训练集样本数和检验集样本数; O_i, $O_{i'}$和 T_i, $T_{i'}$ 分别为训练样本 i 和检验样本 i' 的网络输出与期望输出.

过拟合不确定关系式(4)与信息论中的测不准关系式(1)在形式上虽然极为类似, 内涵上却

有所不同. 其内涵与计算不确定性原理更接近. 数值实验和理论分析表明, 数值计算中存在不确定性原理[17,18], 由数值方法的不准确所导致的不确定性Δe(方法误差)与由计算机的有限精度所造成的不确定性Δr(舍入误差)之间满足一个不确定关系式$\Delta e \cdot \Delta r \geqslant \eta$, 当计算机精度有限时, η为正数. 与此类似, 在 BP 网络过拟合时, 由于训练样本集总是有限而导致的对训练集以外的样本的泛化误差(测试误差)$|\overline{\Delta y}/y|$与计算不确定性原理中由于数值方法不准确导致的方法误差Δe 相对应; 采用函数逼近的 BP 网络由于逼近精度有限造成的训练误差(拟合误差)$|\overline{\Delta z}/z|$与计算不确定性原理中由于计算机有限精度所造成的舍入误差Δr 相对应. 因而泛化误差与训练误差之间也应满足对应的一个不确定关系式$|\overline{\Delta y}/y| \cdot |\overline{\Delta z}/z| \geqslant C$, 当逼近精度有限时, C 亦为正数. 此即表明: BP 网络过拟合出现后, 随着训练继续进行, 泛化误差与训练误差呈反向变化. 因此, 过拟合不确定关系式(4)是不确定性原理在 BP 网络数值模拟过程中的表现.

2 确定过拟合参数 q 的数值模拟实验

本实验构建具有输入层、隐层和输出层的 3 层 BP 网络. 为简便, 输出层只考虑一个输出节点 $k = 1$. 数值模拟实验过程中, 为使网络达到给定的泛化能力, 数值模拟实验必须依据网络结构设计最简原则: 对于隐节点数 H 较多的规模大的网络, 需要的样本数较多, 因而复相关系数 R 较小; 反之, 对隐节点数 H 较少的规模小的网络, 样本数相对亦较少, 因而复相关系数 R 较大. 故无论是规模大或小的网络, 若要求网络有一定的泛化能力, 满足不确定关系式(4)右端的信息量 $S = RH$ 差异不明显. 因此, 在确定过拟合参数 q 的数值模拟实验中, 为了适当减少计算量, 又不会失去一般性, 将网络输入节点数 n 限制在 3~15 范围内, 隐节点数限制在 3~30 范围内比较适合, 已能满足多数实际问题的需要. 训练样本数和检验样本数分别在 m = 16~128 个和 m' = 8~64 个范围内随机选取, 但每次数值实验的检验样本数 m'约占训练样本数 m 的 $\frac{1}{5} \sim \frac{1}{2}$ 为好. 一共模拟了 12 种不同类型函数数值实验, 模拟的 12 类函数的类型如下: 1) 线性函数: $y=\sum_{i=1}^{n} a_i x_i$; 2) 对数函数: $y=\sum_{i=1}^{n} a_i \ln(x_i+i)$; 3) 指数函数: $y=\sum_{i=1}^{n} a_i \mathrm{e}^{x_i}$; 4) 幂函数: $y=\sum_{i=1}^{n} a_i x_i^i$; 5) 正弦函数: $y=\sum_{i=1}^{n} a_i \sin(ix_i)$; 6) 余弦函数: $y=\sum_{i=1}^{n} a_i \cos(ix_i)$; 7) 正切函数: $y=\sum_{i=1}^{n} a_i \mathrm{tg}(ix_i)$; 8) 反正弦函数: $y=\sum_{i=1}^{n} a_i \arcsin(ix_i)$; 9) 反余弦函数: $y=\sum_{i=1}^{n} a_i \arccos(ix_i)$; 10) 反正切函数: $y=\sum_{i=1}^{n} a_i \mathrm{arctg}(ix_i)$; 11) 双曲正、余弦函数: $y=\sum_{i=1}^{n} a_i (\mathrm{e}^{ix_i} \mp \mathrm{e}^{-ix_i})/2$; 12) 复杂函数. 赋予各类函数中的系数 a_i 为 0 ~ $\pm$100 内的随机数; x_i 为样本的第 i 个因子值, 其赋值范围为 0 ~ $\pm$100/10 之间满足函数定义域的随机数.

数值模拟实验过程中, 采用各类模拟函数定义域内因子随机选取值的归一化值作为网络的因子输入值; 训练样本的期望输出值是各类模拟函数计算值的归一化值. 对上述 12 类函数, 分别进行了数百组不同的过拟合数值实验, 得出一致结果: 当过拟合现象出现时, 关系式(4)中的过拟合参数 q 的取值范围一般为 $q \in [\,7\times10^{-3}, 7\times10^{-2}\,]$, 只有极少数例外. 此关系式说明:

构造一定结构的 3 层 BP 网络, 对某给定样本集进行训练, 当训练进行到一定阶段出现过拟合现象后, 若训练继续进行, 虽然可以使训练样本集的拟合误差$|\overline{\Delta z}/z|$进一步减小, 提高了网络学习能力. 但因关系式(4)右端的复相关系数 R, 网络隐节点数 H 确定, 过拟合参数 q 变化范围也确定, 因子 $\log_2(1+M/N)$ 变化很小, 因此, 可认为关系式右端近似保持不变. 为保持关系式(4)成立, 当左端训练样本集的训练相对误差$|\overline{\Delta z}/z|$进一步减小的同时, 检验样本集的测试相对误差$|\overline{\Delta y}/y|$必然增大, 即泛化能力降低, 这正是过拟合现象出现的必然结果. 因此, 关系式(4)可视作 BP 网络过拟合现象出现时, 学习能力与泛化能力之间应满足的一般不确定关系式.

3　不确定关系式的几点讨论

由过拟合不确定关系式(4)可以得出:

1) 由于训练样本集总是有限的, 因此, 网络对训练集以外的样本的泛化误差是不可避免的; 一定结构的网络对函数的逼近能力(训练误差)也是有限的. 而过拟合不确定关系式(4)给 BP 网络的泛化能力与学习能力之间的关系加上了确定的限制, 虽然这个限制的大小常常与模拟对象有关, 但这个限制的存在是固有的, 是不能通过改进网络结构设计或改善样本资料(样本数、样本质量和样本代表性等)加以克服的, 而改进网络结构设计(指调整隐节点数 H)或改善样本资料(表现为函数的复杂性 R 大小发生改变)只能使泛化能力与学习能力之间逐渐达到较佳匹配关系. 这个较佳匹配关系即是不确定关系式(4)取等号时出现. 上述二者之间的关系与文献[18,19]中得出的关系是相容的.

2) 当网络结构给定, 即隐节点数 H 保持不变, 若样本数 m 增加, 则复相关系数 R 变小, 此时, (4)式右端不确定性程度减小, 网络泛化能力增强; 反之, 若样本数 m 减少, 则不确定性程度增大, 网络泛化能力减弱.

3) 当问题给定(因子数 n、样本数 m 和复相关系数 R 保持一定), 若减小网络结构, 即隐节点数 H 减少, 此时(4)式右端不确定性程度亦减小, 网络泛化能力增强; 反之, 若增大网络结构, 则不确定性程度增大, 网络泛化能力减弱.

4) 若样本数 m 减少, 则复相关系数 R 增大, 且若隐节点数 H 增多, 此时(4)式右端不确定性程度增大, 网络泛化能力减小. 这说明网络设计能力超过样本数需求, 故过拟合现象显著; 反之, 若样本数 m 增多, 则复相关系 R 减小, 若隐节点数 H 减少, 此时(4)式右端不确定性程度减弱, 网络泛化能力增强. 这说明相对于较小的网络结构, 若样本数目已足够多, 则不易出现过拟合现象.

5) 若逼近函数形式不变, 当保持隐节点数 H 不变, 样本数 m 不变, 若因子数 n 增加, 则复相关系数 R 增大, 此时(4)式右端不确定性程度增大, 网络泛化能力减弱; 反之, 若因子数 n 减少, 则不确定性程度减小, 网络泛化能力增强.

上述 2) ~ 5)结果均遵循网络结构设计最简原则, 与文献[2, 10, 11]中指出的结果一致, 并与数值模拟实验实际观察到的结果相符合.

4　根据逼近误差要求和样本复杂性选取网络隐节点数

关系式(4)可用于指导网络结构的设计. 对于一个训练好的网络, 若要求训练样本集的平

均训练相对误差(逼近误差)为ε_1，则 N/M 可近似用ε_1代替.

数值模拟实验过程中，不断观察检验样本集的测试相对误差变化，当测试相对误差达最小时，网络的泛化能力与学习能力有最佳的匹配. 在训练相对误差$\varepsilon_1 \leqslant 0.5$ 情况下($\varepsilon_1 > 0.5$ 已没有实际意义)，此时的训练样本集的训练相对误差 $\varepsilon_1 = |\overline{\Delta z}/z|$ 与检验样本集的测试相对误差 $\varepsilon_2 = |\overline{\Delta y}/y|$之间近似地可用以下公式相联系:

$$\varepsilon_2 = (\varepsilon_1/2)^{\frac{1}{2}} \quad (\varepsilon_1 \leqslant 0.50). \tag{9}$$

记过拟合参数 q 的取值范围为 $q \in [q_1, q_2] = [7\times10^{-3}, 7\times10^{-2}]$，将(9)式代入(4)式左端，若(4)式右端的过拟合参数 q 值按 0.618 优选法选取其第 1 个优点值 $q' = (q_2 - q_1)\times 0.618 + q_1 = 4.59\times10^{-2}$，则由不确定关系式(4)得

$$H \leqslant \frac{30.8}{R}\varepsilon_1^{\frac{3}{2}}\log_2\left(1+\frac{1}{\varepsilon_1}\right), \tag{10}$$

按四舍五入法取整，并取“ = ”得

$$H^* = \left[\frac{30.8}{R}\varepsilon_1^{\frac{3}{2}}\log_2\left(1+\frac{1}{\varepsilon_1}\right)\right]. \tag{11}$$

(11)式表明: 若希望构造的网络有好的泛化能力，当过拟合参数 q 取定后，其隐节点数的选取由样本的复杂性 R，问题要求的训练样本的训练相对误差ε_1的取值确定. 而复杂性又与问题规模(样本数 m 和因子数 n)和函数的性质有关. 可见，样本(模拟函数)愈复杂，R 愈小，所需隐节点数愈多，即网络结构愈复杂；要求的训练误差 ε_1 愈小，所需隐节点数愈少，即网络结构愈简单. 因此，若应用 BP 网络对复杂样本进行训练，而要求的训练误差又要小，则构建一个具有较好泛化能力的网络的隐节点数的选取应取决于(11)式右边的 R 和ε_1的协调.

5　用选取的隐节点数构建 BP 网络的合理性验证

为了验证由(11)式所确定的隐节点数 H^*构建的 BP 网络是否合理，用(11)式确定的隐节点数 H^*构建 BP 网络，对 5 类函数进行了数值模拟实验. 其中的第 5 类复杂函数是由多种不同的初等函数任意组合构成的. 例如表 1 中的第 1 个复杂函数可如下构成:

$$y = 19\,\mathrm{th}x_1 - 247\arcsin\frac{2x_2}{1+x_2^2} + 172\sin^2 x_3\cos x_4^2 + 0.91\,\mathrm{ch}\sqrt[5]{x_5^2+x_6^2} + 74\lg(1001+x_7).$$

根据预先指定的训练误差ε_1，及计算出的函数的复相关系数 R 值，代入(11)式，计算得到所需构造的 BP 网络的隐节点数 H^*如表 1 所示.

以隐节点数 H^*构建 BP 网络，进行过拟合参数 q 的数值模拟实验，不同实验中的样本因子数 n，训练样本数 m，检验样本数 m'，计算出的复相关系数 R 及过拟合数值模拟实验结果亦见表 1. 各类函数不同过拟合实验计算确定的过拟合参数 q 的值如表 1 最右一列所示. 的确，其 q 值分布范围均在[7×10^{-3}，7×10^{-2}]内，表明用(11)式确定出的隐节点数 H^*构造 BP 网络既能大致满足给定训练样本的训练误差要求，又能使网络有较佳的泛化能力. 因此，由(11)式所确定的隐节点数构造 BP 网络是一种合理的选择.

文献[20]中给出了隐节点数选择的参考公式: 1) $k < \sum_{i=1}^{n} C_i^{n_1}$；2) $n_1 = \sqrt{n+m} + a$　(a = 1 ~ 10

表 1　几类函数的隐节点数 H^* 选取的数值模拟实验及过拟合验证结果

函数类型	ε_1	R	H^*	n	m	m'	$\left\lvert\frac{\overline{\Delta z}}{z}\right\rvert$	$\left\lvert\frac{\overline{\Delta y}}{y}\right\rvert$	M/N	$q\ (1\times10^{-2})$
线性函数	0.05	1.000	2	6	104	13	0.0477	0.1933	19.206	3.98
	0.08	1.000	3	3	58	28	0.0894	0.1953	8.043	3.66
	0.10	1.000	3	7	27	10	0.0833	0.2197	9.859	4.17
	0.13	1.000	5	6	22	14	0.1343	0.2361	7.890	3.89
	0.15	1.000	5	5	51	34	0.1795	0.2416	5.807	4.75
对数函数	0.05	0.9051	2	3	56	23	0.0501	0.1258	20.862	3.07
	0.08	0.8435	3	4	124	48	0.0921	0.2098	10.588	5.31
	0.10	0.9012	4	5	49	30	0.1071	0.2758	8.421	5.22
	0.15	0.8745	6	4	79	61	0.1491	0.2698	6.996	4.49
	0.20	0.8618	8	5	24	13	0.2075	0.3216	6.691	5.64
指数函数	0.07	0.5585	4	7	30	16	0.0734	0.2143	16.357	5.70
	0.10	0.5075	7	13	97	13	0.1161	0.2870	9.247	6.24
	0.13	0.4630	10	11	34	11	0.1327	0.2593	6.086	4.05
	0.15	0.4886	11	9	25	15	0.1780	0.2740	8.158	5.72
	0.20	0.4534	16	15	45	26	0.2297	0.3445	5.151	5.82
幂函数	0.08	0.8805	3	3	100	44	0.0709	0.1877	12.647	3.73
	0.10	0.8320	4	7	63	30	0.1206	0.2339	7.043	5.03
	0.13	0.8048	6	5	76	37	0.1499	0.2461	6.030	4.24
	0.15	0.7452	7	9	102	20	0.1585	0.2996	5.456	4.79
	0.20	0.7407	10	13	59	22	0.2374	0.3011	5.950	5.33
复杂函数	0.08	0.3005	9	7	65	43	0.0905	0.2983	9.595	6.67
	0.10	0.3232	10	12	24	10	0.1093	0.2799	7.669	5.74
	0.13	0.3678	12	5	78	35	0.1323	0.2674	6.464	4.48
	0.15	0.3869	14	8	56	45	0.1979	0.2171	4.995	3.99
	0.20	0.3465	21	15	48	21	0.1823	0.3717	5.426	4.87

之间常数); 3) $n_1 = \log_2 n$. 以上诸式中, n_1 为隐节点数, n 为输入节点数, m 为输出节点数, k 为样本数. 由上述3个公式确定的隐节点数 n_1 与函数的复杂性和训练精度要求没有联系, 且由上述3个公式确定的隐节点数 n_1 差异很大.

本文给出的隐节点数 H^* 的计算公式(11)是由不确定关系式(4)导出的. 而(4)式是基于网络结构的复杂性与问题规模相匹配的网络结构设计的最简原则, 因而它的导出有较充分的理论依据. 只要给定问题要求的训练误差 ε_1 和由(2)式计算出复相关系数 R, 就可由(11)式确定有一定的泛化能力的 BP 网络所需选择的隐节点数 H^*. 由于复相关系数 R 既反映了问题性质本身的复杂性, 又隐含了问题规模(样本数 m 和因子数 n)所具有的复杂性, 因此, 由(11)式确定的隐节点数同时考虑了多种因素的影响, 是一种较合理的选择.

6　改进 BP 网络泛化能力的最佳停止训练法

关系式(4)还可用于指导 BP 网络训练, 改进泛化能力. 判别方法如下: 一般说来, BP 网络训练过程中, 刚开始出现过拟合时, 训练样本集的平均训练相对误差 $\varepsilon_1 = |\overline{\Delta z}/z|$ 与对检验样本

集的平均测试相对误差 $\varepsilon_2 = |\overline{\Delta y} / y|$ 近似地通过(9)式相联系. 因此, 由(4)式有

$$\frac{1}{\sqrt{2}}\varepsilon_1^{\frac{3}{2}} \geqslant \frac{RHq}{2\log_2(1+M/N)},$$

故有

$$q \leqslant \sqrt{2}\varepsilon_1^{\frac{3}{2}} \frac{\log_2(1+M/N)}{RH}. \tag{12}$$

由于过拟合参数 $q \in [7\times10^{-3}, 7\times10^{-2}]$, 故上式右端限制范围为

$$1\times10^{-3} \leqslant \frac{\sqrt{2}\varepsilon_1^{\frac{3}{2}} \log_2(1+M/N)}{7RH} \leqslant 1\times10^{-2}. \tag{13}$$

BP 网络用于某一具体问题建模, R 和 H 是确定的; 在训练过程中, 样本集每训练完一遍, 可以计算出 ε_1 和 M/N 值. 因此, 只要在训练过程中, 不断观察 $q' = \frac{\sqrt{2}\varepsilon_1^{\frac{3}{2}} \log_2(1+M/N)}{7RH}$ 值的变化. 随着训练的进行, ε_1 不断减小, 从而 q' 亦不断减小, 若在 BP 网络达到指定精度停止训练前, 总有 $q' > 1\times10^{-2}$, 则认为训练过程中, 并未出现过拟合. 若在停止训练前, 已有 $q' \leqslant 1\times10^{-2}$, 甚至 $q' \leqslant 1\times10^{-3}$, 则可认为训练过程中已出现过拟合, 尽管训练还未达到指定精度, 也必须停止训练. 可见关系式(4)对实际问题的 BP 网络建模训练过程中防止出现过拟合, 改进泛化能力具有指导意义.

7 结论

1) 基于计算不确定性原理和神经网络结构设计最简原则建立的训练样本集的训练相对误差与检验样本集的测试相对误差之间满足的不确定关系式(4), 直接、明确、简单地揭示了 BP 网络过拟合时的学习能力与泛化能力之间满足的定量关系式. 通过大量数值模拟实验, 确定了关系式(4)中的过拟合参数 q 的取值范围为 $q \in [7\times10^{-3}, 7\times10^{-2}]$.

2) 过拟合不确定关系式(4)给出了 BP 网络对给定问题模拟能力(逼近精度)的确定的上限, 超过这个上限, 函数的模拟能力愈强(训练误差愈小), 则泛化能力愈弱(泛化误差愈大). 并指出了 BP 网络对给定样本集训练过程中, 为达到最佳泛化能力的实现途径.

3) 其他前向神经网络过拟合时也应有与关系式(4)相似的形式, 只是不同的前向网络关系式右端某些字母代表的具体含义和确定出的参数的取值范围有所不同. 因此, 关系式(4)形式具有普遍意义.

4) 基于(11)式的 BP 网络结构的隐节点数的选取是在满足问题给定精度要求下并使网络具有较佳泛化能力的合理选择, 其合理性通过数值实验得到验证. 公式形式简洁, 计算简单. 对一个给定的训练样本集, 只要计算出它的复相关系数 R 和给定模型精度要求 ε_1, 即可易于确定隐节点数 H^*.

5) 关系式(4)建立过程中, 只考虑了网络结构、函数复杂性和问题规模对学习能力和泛化能力的影响, 并未考虑样本质量、样本代表性、初始权值和先验知识等对泛化能力的影响; 此外, 复相关系数 R 并不能完全表征函数(问题)的复杂性, 隐节点数的确定也是一个复杂而困难的问题, 本文给出的确定隐节点数公式(11)也只是一种尝试. 这些问题有待于进一步深入探索.

参　考　文　献

1 Baum E B, Haussler D. What Size Net Gives Valid Generalization? San Mateo, NIPSI, 1989. 81~90
2 Moody J E. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning system. San Mateo, NIPS 4, 1992. 847~854
3 Barron A R. Approximation and estimation bounds for artificial neural networks. Machine Learning, 1994, 14: 115~133
4 Geman S. Neural networks and bias/variance dilemma. Neural Computation, 1992, 4: 1~58
5 Cataltepe Z, Abu-mostafa Y S, Magdon-Ismail M. No free lunch for earlystopping. Neural Computation, 1999, 11: 995~1009
6 Amari S, Murata N, Muller K R, et al. A symptotic statistical theory of overtraining and cross-validation. IEEE Trans Neural Networks, 1997, 8(5): 985~996
7 Baldi P. Temporal evolution of generalization during learning in linear networks. Neural Computation, 1991, 3: 589~603
8 Partridge D. Network generalization differences quantified. Neural Networks, 1996, 9(2): 263~271
9 张鸿滨. 训练多层网络的样本数问题. 自动化学报, 1993, 19(1): 71~77
10 魏海坤, 徐嗣鑫, 宋文忠. 神经网络的泛化理论和泛化方法. 自动化学报, 2001, 27(6): 806~815
11 阎平凡. 人工神经网络的容量、学习与计算复杂性. 电子学报, 1995, 23(4): 63~67
12 王　晖, 何新贵. BP 网络泛化能力改进研究. 系统工程与电子技术, 2001, 23(3): 85~87
13 江学军, 唐焕文. 前馈神经网络泛化性能力的系统分析. 系统工程理论与实践, 2000, 20(8): 38~40
14 彭汉川, 甘　强, 韦　钰. 提高前馈神经网络推广能力的若干实际方法. 计算机工程与应用, 1999, 1: 47~48
15 Zha Youliang. Information uncertainty principle. Chinese Science Bulletin, 1989, 34(1): 86~87
16 施　能. 气象统计预报中的多元分析方法. 北京: 气象出版社, 1992. 30~33
17 李建平. 算不准原理及其意义与启示. 中国科学院院刊, 2000, 15(6): 428~430
18 李建平, 曾庆存, 丑纪范. 非线性常微分方程的计算不确定性原理——Ⅱ. 理论分析. 中国科学, E 辑, 2000, 30(6): 550~567
19 阎平凡. 不确定原则和神经网络用于信号恢复. 信号处理, 1991, 7(2): 71~76
20 张立明. 人工神经网络的模型及其应用. 上海: 复旦大学出版社, 1993