

梯度算法下RBF网的参数变化动态

魏海坤, 李 奇, 宋文忠

(东南大学 自动化学院, 江苏 南京 210096)

摘要: 分析神经网络学习过程中各参数的变化动态, 对理解网络的动力学行为, 改进网络的结构和性能等具有积极意义. 本文讨论了用梯度算法优化误差平方和损失函数时RBF网隐节点参数的变化动态, 即算法收敛后各隐节点参数的可能取值. 主要结论包括: 如果算法收敛后损失函数不为零, 则各隐节点将位于样本输入的加权聚类中心; 如果损失函数为零, 则网络中的冗余隐节点将出现萎缩、衰减、外移或重合现象. 进一步的试验发现, 对结构过大的RBF网, 冗余隐节点的萎缩、外移、衰减和重合是频繁出现的现象.

关键词: 梯度算法; RBF网; 学习动态; 神经网络; 泛化能力

中图分类号: TP183 **文献标识码:** A

Gradient learning dynamics of radial basis function networks

WEI Hai-kun, LI Qi, SONG Wen-zhong

(School of Automation, Southeast University, Nanjing Jiangsu 210096, China)

Abstract: To understand the dynamic behavior and improve the structure and performance of neural networks, it is very important to investigate their parameter changing dynamics during the learning. For radial basis function (RBF) networks using gradient descent method to minimize the least squares error cost function, this paper discusses the learning dynamics of the hidden unit parameters, i.e., their possible values after learning. It is proved that if the cost function is not zero after the algorithm converges, then all hidden units will move to the weighted cluster centers of sample inputs. If cost function is zero, then hidden units will have shrinking, eliminating, out-moving and overlapping happened to those redundant units. Further simulation shows that such phenomena occur frequently in oversized RBF networks.

Key words: gradient method; RBF network; learning dynamics; neural networks; generalization ability

1 引言(Introduction)

前馈神经网络, 包括多层感知器(MLPs)、RBF网等, 在实际中能否成功应用, 取决于网络的泛化能力, 即训练后的神经网络对工作样本或测试样本作出正确反应的能力. 神经网络的泛化理论和泛化方法是神经网络理论和应用研究的重要内容^[1]. 影响神经网络泛化能力的因素很多, 包括训练样本的数量和质量、网络的结构复杂性、对所学习的目标规则的先验知识、网络的初始权值、训练时间等. 上述因素中, 神经网络的结构复杂性对网络泛化能力影响极大. 对于给定的目标规则和一组训练样本, 如果网络结构过于简单, 则模型的描述能力不足, 将无法实现对目标规则的学习, 从而出现所谓的欠拟合; 如果神经网络结构过于复杂, 则模型的描述能力过剩, 将出现所谓的过拟合. 因此, 为了尽可能提高神经网络的泛化能力, 通常需要设计满足精度要求的结构

较小的神经网络^[2].

RBF网同样具有任意逼近能力, 而且比MLPs结构更简单, 学习速度也更快, 因此引起了许多人的重视. RBF网的学习方法很多, 其中许多方法都使用梯度算法调节网络内的所有或部分参数^[3~6]. 此时对RBF网来说, 学习过程中对各隐节点中心、扩展常数和输出权值的调整, 最终体现为各隐节点中心的位置、隐节点的影响范围及其输出权值的变化. 分析这些隐节点参数的变化动态, 对理解网络的动力学行为, 改进网络的泛化能力, 指导网络结构设计等具有积极意义. 本文研究了用梯度算法最小化误差平方和损失函数(即最小二乘误差)时RBF网参数的变化动态.

本文第2节介绍有关的一些基本定义; 在第3节, 就算法收敛时损失函数为零与否两种情况, 分别分析了RBF网隐节点参数的取值; 第4节验证了只要隐

节点取特定初始参数, 就可以出现相应的学习动态, 并通过仿真说明了冗余隐节点的萎缩、外移、衰减和重合是频繁出现的现象; 最后讨论了进一步的工作.

2 定义(Definition)

考虑一个 $m - h - 1$ 结构的RBF网, 其中 m 和 h 分别为网络的输入节点和隐节点数. 假定第 i 个隐节点(简称为隐节点 i) 的激活函数为 $\phi_i(\mathbf{x})$, 其形式采用最常用的Gaussian型径向基函数:

$$\phi_i(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_i\|^2}{r_i^2}\right), \quad (1)$$

其中: $\mathbf{x} \in \mathbb{R}^m$ 为RBF网输入, $\mathbf{c}_i \in \mathbb{R}^m$ 为隐节点 i 的中心, r_i 为相应的扩展常数(也称为宽度).

用三元组 $\theta_i = \{\mathbf{c}_i, w_i, r_i\}$ 表示隐节点 i 的所有参数, 其中 w_i 为其输出权值. 于是含 h 个隐节点的RBF网的所有参数可表示为 $\xi^{(h)} = \{\theta_1, \theta_2, \dots, \theta_h\}$. 显然, 此时的RBF网可由 $\xi^{(h)}$ 唯一标识. 由此得到RBF网所实现的输入输出映射为

$$f(\mathbf{x}, \xi^{(h)}) = \sum_{i=1}^h w_i \phi_i(\mathbf{x}). \quad (2)$$

假定训练样本集为 $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, 其中 $\mathbf{x}_i \in \mathbb{R}^m$ 为样本输入, $y_i \in \mathbb{R}$ 为相应的样本输出, N 为训练样本数. 于是样本输出矢量为 $\mathbf{Y} = [y_1, y_2, \dots, y_N]^T$. 令隐节点 i 对所有样本输入的响应矢量为 $\mathbf{o}_i = [o_{i1}, o_{i2}, \dots, o_{iN}]^T$, 其中 $o_{ij} = \exp\left(-\frac{\|\mathbf{x}_j - \mathbf{c}_i\|^2}{r_i^2}\right)$ 为输入 \mathbf{x}_j 时隐节点 i 的输出. 另外, 假定所有样本输入组成的有效样本输入区域为 D , 显然 $D \subset \mathbb{R}^m$.

RBF网学习的核心内容, 是确定一组最优的网络参数 $\xi^{*(h)}$, 使得

$$y_i = f(\mathbf{x}_i, \xi^{*(h)}). \quad (3)$$

采用以下的误差平方和损失函数:

$$E(\xi^{(h)}) = \sum_{i=1}^N (y_i - f(\mathbf{x}_i, \xi^{(h)}))^2, \quad (4)$$

上式对应的矢量形式为 $E(\xi^{(h)}) = \mathbf{e}_h^T \mathbf{e}_h$, 其中 $\mathbf{e}_h = \mathbf{Y} - \sum_{i=1}^h w_i \mathbf{o}_i$. 于是当网络对训练样本实现完全内插时, 其参数 $\xi^{*(h)}$ 必然满足 $E(\xi^{*(h)}) = 0$. 此时根据下面的Michelli定理^[7], 对于训练样本集 S , 总可以找到一RBF网 $\xi^{*(h)}$, 使之能内插 S 中的所有样本.

Michelli定理: 假定 ϕ 是区间 $[0, \infty)$ 上的连续函数, 且在区间 $(0, \infty)$ 上恒为正. 如果其一阶导函数在区间 $(0, \infty)$ 上完全单调且不为常数, 则对 \mathbb{R}^m 中

的 N 个不同的点 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$, 有

$$(-1)^N \det \phi(\|\mathbf{x}_i - \mathbf{x}_j\|^2) > 0.$$

Micchelli定理只是说明对任意 N 个训练样本, 总可以用 N 个隐节点的RBF网实现内插. 当样本数较多时, 这显然不是一种好的逼近方式. 因此较常用的方法是暂时固定隐节点数 h , 然后以式(4)为损失函数, 并用梯度法训练各隐节点参数. 但是, 由于并不知道什么样的隐节点数是合理的, 因此常选取一个足够大的初始网络. 于是算法收敛时网络参数既可能收敛到损失函数的某个局部最小点 ($E(\xi^{(h)}) \neq 0$), 也可能收敛到损失函数的某个全局最小点 ($E(\xi^{(h)}) = 0$). 本文将对这两种情况, 讨论梯度算法下隐节点参数的取值.

为便于后面描述, 再引入以下概念:

隐节点重合: 算法收敛时两个隐节点具有相同的节点中心和扩展常数的现象.

隐节点外移: 而算法收敛时隐节点 i 的节点中心已位于有效样本输入区域 D 外, 且其扩展常数也已经不足以影响 D 内的样本, 导致 $\mathbf{o}_i = 0$ 的现象.

隐节点衰减: 算法收敛后隐节点 i 的输出权值 $w_i = 0$ 的现象.

隐节点萎缩: 算法收敛隐节点 i 的节点中心虽位于有效样本输入区域 D 内, 且其扩展常数较小, 已经不足以影响附近的样本, 导致 $\mathbf{o}_i = 0$ 的现象.

3 主要结论(Main results)

3.1 学习后损失函数不为零(Cost function is not zero after learning)

虽然讨论的是RBF网学习时所有隐节点参数的变化情况, 但不失一般性, 可以先考虑最后一个隐节点(即隐节点 h) 参数, 即 w_h, \mathbf{c}_h, r_h 的取值.

令 $\mathbf{e}_{h-1,j} = y_j - \sum_{i=1}^{h-1} w_i o_{ij}$, 则 $\mathbf{e}_{h-1} = \mathbf{Y} - \sum_{i=1}^{h-1} w_i \mathbf{o}_i$ 表示前 $h-1$ 个隐节点对目标输出的逼近残差矢量, 于是 $\mathbf{e}_h = \mathbf{e}_{h-1} - w_h \mathbf{o}_h$, 且有

$$E(\xi^{(h)}) = \mathbf{e}_{h-1}^T \mathbf{e}_{h-1} - 2w_h \mathbf{e}_{h-1}^T \mathbf{o}_h + w_h^2 \mathbf{o}_h^T \mathbf{o}_h. \quad (5)$$

求解 $E(\xi^{(h)})$ 对 \mathbf{c}_h, r_h 和 w_h 的偏导数, 得到:

$$\frac{\partial E}{\partial \mathbf{c}_h} = -\frac{4w_h}{r_h^2} \sum_{i=1}^N e_{h,i} o_{hi} (\mathbf{x}_i - \mathbf{c}_h), \quad (6)$$

$$\frac{\partial E}{\partial r_h} = -\frac{4w_h}{r_h^3} \sum_{i=1}^N e_{h,i} o_{hi} \|\mathbf{x}_i - \mathbf{c}_h\|^2, \quad (7)$$

$$\frac{\partial E}{\partial w_h} = -2(\mathbf{e}_{h-1} - w_h \mathbf{o}_h)^T \mathbf{o}_h. \quad (8)$$

当算法收敛时式(6)~(8)为零, 此时可解得

$$\mathbf{c}_h = \frac{\sum_{i=1}^N e_{h,i} o_{hi} \mathbf{x}_i}{\sum_{i=1}^N e_{h,i} o_{hi}}, \quad (9)$$

$$\sum_{i=1}^N e_{h,i} o_{hi} \|\mathbf{x}_i - \mathbf{c}_h\|^2 = 0, \quad (10)$$

$$w_h = \frac{\mathbf{e}_{h-1}^T \mathbf{o}_h}{\mathbf{o}_h^T \mathbf{o}_h}. \quad (11)$$

N 个样本输入的聚类中心为 $\mathbf{C} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$. 如果令 $k_i = \frac{e_{h,i} o_{hi}}{\sum_{j=1}^N e_{h,j} o_{hj}}$, 则式(9)为:

$$\mathbf{c}_h = \sum_{i=1}^N k_i \mathbf{x}_i. \quad (12)$$

此时 \mathbf{c}_h 可看作 N 个样本输入“加权聚类中心”. 从加权系数 k_i 的表达式可见: 如果某样本输入 \mathbf{x}_i 不能激活隐节点 (o_{hi} 值较小), 或者当前神经网络对该样本的误差较小 (e_{hi} 较小), 则该 \mathbf{x}_i 对 \mathbf{c}_h 的影响就较小. 隐节点 h 的中心最终移向那些未被其它隐节点激活且偏差较大的样本输入. 这也与笔者的直觉是一致的.

式(10)是算法收敛时 r_h 必须满足的约束条件. 当 r_h 调整至最佳时 \mathbf{o}_h 和 \mathbf{e}_h 应该是互相垂直的, 即 $\mathbf{o}_h^T \mathbf{e}_h = 0$. 但这里, $o'_{h,i} = d_i o_{h,i}$, 其中 $d_i = \|\mathbf{x}_i - \mathbf{c}_i\|^2$, 说明 r_h 的作用是隐节点 h 的加权输出 $\mathbf{o}'_h = [o'_{h,1}, o'_{h,2}, \dots, o'_{h,N}]^T$ 与 \mathbf{e}_h 互相垂直. 因此权重系数 d_i 的作用是惩罚 $\|\mathbf{x} - \mathbf{c}_i\| = 0$ 附近样本对应的隐节点输出, 因为在 $\|\mathbf{x} - \mathbf{c}_i\| = 0$ 附近的样本, 其相应的隐节点输出受 r_h 变化的影响较小.

式(11)的几何意义是很明显的, 它是使 \mathbf{e}_{h-1} 和 \mathbf{o}_h 间距离达最小时系数 w_i 的取值.

于是有以下结论:

结论 1 如果算法收敛时损失函数不为零, 则各隐节点将移向所有样本输入的加权聚类中心.

3.2 学习后损失函数为零 (Cost function is zero after learning)

考虑网络少于 h 个隐节点就能够对训练样本实现内插的情况. 此时如果算法收敛时 $E(\boldsymbol{\xi}_{(h)}) = 0$, 那么各隐节点参数将收敛到什么值? 这里考虑只有一个冗余节点, 且该隐节点为 h 的情况.

如果网络的前 $h-1$ 个隐节点已经能够内插训练样本, 则当隐节点 h 的输出权值 w_h 很小; 或其节点中心 \mathbf{c}_h 远离样本有效输入区域; 或 \mathbf{c}_h 在样本有效输入区域内但远离任何一个样本输入, 且 r_h 足够小, 则任何一个训练样本都将无法激活该隐节点, 此时该隐

节点将是冗余的.

另外, 如果网络的前 $h-1$ 个隐节点已经能够对训练样本实现内插, 而隐节点 h 与某个隐节点 $j < h$ 重合, 则隐节点 h 也将是冗余的.

综上所述, 有下面的结论:

结论 2 如果 $\boldsymbol{\xi}_{(h-1)} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_{h-1}\}$ 能对训练样本实现内插, 则当 $\boldsymbol{\xi}_{(h)}^* = \{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \dots, \boldsymbol{\theta}_h^*\}$ 满足以下两个条件中的一个时, 也能对训练样本实现内插: 1) 隐节点 h 的参数 $\boldsymbol{\theta}_h$ 满足 $w_h \approx 0$; 或 $\mathbf{c}_h = \infty$; 或 $\mathbf{c}_h \neq \mathbf{c}_i$, 且 $r_h \approx 0$; 而其它隐节点与 $\boldsymbol{\xi}_{(h)}$ 的隐节点完全重合, 即 $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^*$, $1 \leq i < h$; 2) 隐节点 h 与某个隐节点 $j < h$ 重合, 且满足 $w_j = w_j^* + w_h^*$, 而 $\boldsymbol{\xi}_{(h)}^*$ 的其它隐节点与 $\boldsymbol{\xi}_{(h-1)}$ 完全重合. 即 $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^*$, $1 \leq i < h$.

结论2说明, 如果网络中存在冗余隐节点, 且当前网络参数在 $\boldsymbol{\xi}_{(h)}^*$ 的某个邻域内, 则在梯度算法作用下, 冗余隐节点将出现重合、外移、萎缩和衰减现象.

4 仿真例 (Simulation)

4.1 隐节点参数变化动态的演示例子 (Demonstration of learning dynamics)

上面讨论的是 RBF 网训练时各隐节点的学习动态. 这里给出了若干隐节点参数的初始条件, 只要满足这些初始条件, 隐节点将出现相应的动态. 产生训练样本的目标函数如下:

$$g(x_1, x_2) = 2.0 \exp\left(-\frac{(x_1 - 0.3)^2 + (x_2 - 0.7)^2}{0.04}\right) + 1.8 \exp\left(-\frac{(x_1 - 0.6)^2 + (x_2 - 0.4)^2}{0.04}\right). \quad (13)$$

共产生 121 个训练样本 $((x_1, x_2), y)$, 其中 (x_1, x_2) 为样本输入, $y = g(x_1, x_2) + N(0, 0.1)$ 为样本输出. $N(0, 0.1)$ 为均值为 0, 方差为 0.1 的高斯噪声, $x_i \in \{0, 0.1, \dots, 1\}$, $i = 1, 2$. 采用标准梯度算法学习网络的 3 个参数, 训练一定次数后结束.

应该指出, 该例子虽然简单, 但可以直观且充分地演示梯度算法下各隐节点参数的变化动态.

4.1.1 隐节点移向聚类 (Hidden units clustering)

网络使用两个隐节点, 它们的初始参数分别为 $\{(0.15, 0.25), 2.5, 0.45\}$, $\{(0.50, 0.25), 1.3, 0.10\}$. 学习率为 0.002, 经 2000 次学习后算法收敛, 此时网络参数变为 $\{(0.3031, 0.7027), 1.9453, 0.2048\}$, $\{(0.6061, 0.3956), 1.8604, 0.1930\}$.

图1是学习过程中两个节点中心的移动轨迹. 以“□”开始的曲线为两个隐节点中心的移动轨迹, 以“○”开始的曲线为每次参数调节时理论计算得到的两个加权聚类中心的轨迹, 环形线为目标函数的等高线. 由图可见, 两个隐节点的最终的加权聚类中心与目标曲线中的隐节点中心值是一致的, 而且两个隐节点最终都移到了相应的加权聚类中心.

4.1.2 隐节点外移(Hidden units out-moving)

网络使用3个隐节点, 它们的初始参数分别为 $\{(0.25, 0.65), 1.9, 0.35\}$, $\{(0.70, 0.45), 1.6, 0.38\}$, $\{(0.05, 0.75), -1.0, 0.40\}$. 算法收敛后前两个隐节点参数为 $\{(0.2921, 0.7025), 1.9483, 0.2091\}$, $\{(0.5987, 0.3945), 1.8407, 0.1991\}$, 第3个节点参数则为 $\{(-0.3188, 0.6420), -0.8085, 0.2377\}$ (见图2). 从图2可以明显看到, 两个隐节点移到了相应的正确位置, 第3个隐节点则远出了样本输入区域(移出部分轨迹未画出), 出现了隐节点外移现象.

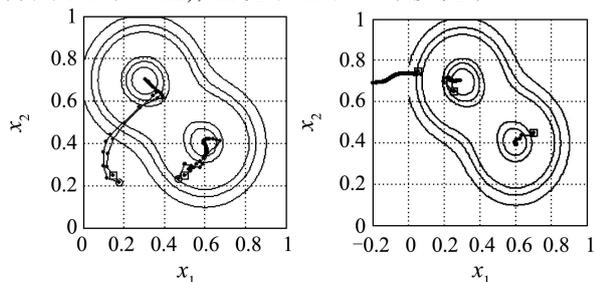


图 1 隐节点聚类 Fig. 1 Units clustering
图 2 隐节点外移 Fig. 2 Unit out-moving

4.1.3 隐节点重合(Hidden units overlapping)

网络使用3个隐节点, 它们的初始参数分别为 $\{(0.10, 0.55), 0.3, 0.08\}$, $\{(0.50, 0.45), 1.6, 0.28\}$, $\{(0.25, 0.85), 0.8, 0.46\}$. 算法收敛后隐节点1和3的参数为 $\{(0.3028, 0.7056), 0.6793, 0.2048\}$, $\{(0.3027, 0.7048), 1.3415, 0.2024\}$, 隐节点2的参数为 $\{(0.6074, 0.3865), 1.7740, 0.2040\}$. 图3(a)是3个隐节点中心的移动轨迹, 图3(b)为两个隐节点宽度的变化曲线. 由图可见, 隐节点1和3的节点中心趋向重合, 同时它们的宽度也趋于相等, 出现了隐节点重合现象. 另外, 它们的输出权值之和为2.0208, 与目标函数中第1个隐节点的输出权值非常接近.

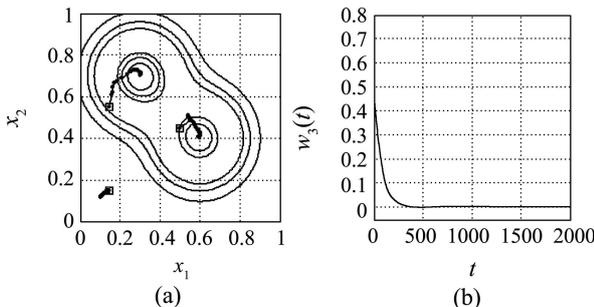


图 3 隐节点重合 Fig. 3 Hidden units overlapping

4.1.4 隐节点衰减(Hidden units decaying)

网络使用3个隐节点, 它们的初始参数分别为 $\{(0.15, 0.55), 1.9, 0.15\}$, $\{(0.50, 0.45), 1.6, 0.28\}$, $\{(0.15, 0.15), 0.6, 0.50\}$. 算法收敛后前两个隐节点参数为 $\{(0.3008, 0.7018), 1.9585, 0.1963\}$, $\{(0.5955,$

$0.4071), 1.8868, 0.1986\}$, 隐节点3的参数则为 $\{(0.1033, 0.1158), -0.0002, 0.4097\}$. 图4(a)是3个隐节点中心的移动轨迹, 图4(b)是隐节点3的输出权值的变化情况. 由图可见, 尽管隐节点3的中心位于样本输入区域内, 且其扩展常数也较大, 但是其输出权值衰减为接近于零, 出现了隐节点衰减现象.

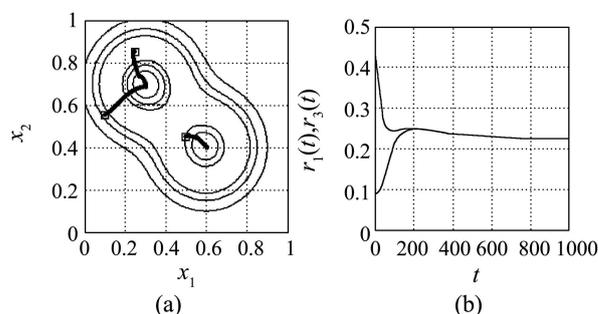


图 4 隐节点衰减 Fig. 4 Hidden units eliminating

4.1.5 隐节点萎缩(Hidden units shrinking)

网络使用3个隐节点, 它们的初始参数分别为 $\{(0.20, 0.70), 2.5, 0.25\}$, $\{(0.70, 0.30), 1.2, 0.15\}$, $\{(0.90, 0.12), 1.0, 0.30\}$. 算法收敛后前两个隐节点参数为 $\{(0.2955, 0.6994), 2.0047, 0.2013\}$, $\{(0.6069, 0.4009), 1.7851, 0.1986\}$, 隐节点3的参数则为 $\{(0.9382, 0.1495), 0.8942, 0.0172\}$.

图5(a)是3个隐节点中心的移动轨迹, 图中“O”处为样本输入. 图5(b)是隐节点3扩展常数的变化情况. 可见该隐节点的最终宽度虽然不为零, 但值较小(0.0172), 致使所有训练样本都无法影响该隐节点. 即便是离其最近的样本输入点(0.9, 0.1), 隐节点输出也已经很小(1.6304×10^{-6}), 其影响已经可以忽略, 出现了萎缩现象. 可以看到, 隐节点萎缩时, 其隐节点中心移到了附近4个样本的“间隙”处.

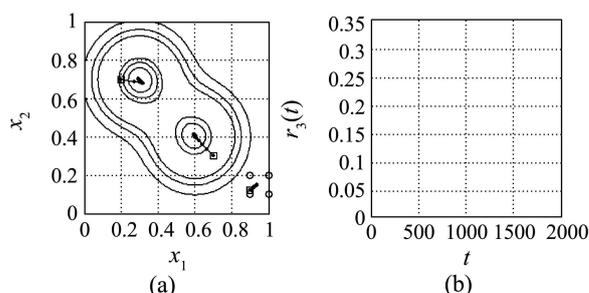


图 5 隐节点萎缩 Fig. 5 Hidden unit shrinking

4.2 冗余隐节点学习动态统计(Statistics of redundant hidden units learning dynamics)

在本例中, 用一个SISO的非线性系统建模的例子对冗余隐节点的各种学习动态的出现情况进行了统计. 目标函数为

$$y(k) = \frac{29}{40} \sin\left(\frac{16u(k-1) + 8y(k-1)}{3 + 4u(k-1)^2 + 4y(k-1)^2}\right) + \frac{2}{10}u(k-1) + \frac{2}{10}y(k-1), \quad (14)$$

其中 $u(k)$ 为输入信号,取 $[-0.1, 0.1]$ 内均匀分布的随机值.令 $y(0) = 0$,按上式产生50个原始样本,然后将这些样本输入归一化到区间 $[0, 1]$ 后作为训练样本.用一个足够大(2-40-1)的RBF网对上式进行建模,网络的两个输入为 $u(k-1)$ 和 $y(k-1)$,输出为 $y(k)$.用梯度法训练网络的节点中心、输出权值和扩展常数,训练时学习参数设置如下:节点中心和扩展常数初值均取 $[0, 0.1]$ 内均匀分布的随机值,初始权值取 $[-0.1, 0.1]$ 内均匀分布的随机值,训练2000后结束,学习率取0.002.

图6(a)显示了某次学习结束后对最近样本激活值(即 $w_i \exp(-\frac{\|\mathbf{x} - \mathbf{c}_i\|^2}{r_i^2})$)小于0.0001的16个隐节点(即未激活隐节点)的中心及其影响范围.图中每个隐节点均表示为两个同心圆:其中固定大小的圆“○”表示节点中心,另一个圆以扩展常数为半径(图中如果扩展常数较小则缩为一个点),表示其影响范围.图中“+”表示训练样本.可见,这些隐节点或者其扩展常数很小,已不足以影响最近的样本,出现了萎缩现象,如图中的隐节点1,2;或者扩展常数虽然较大,但由于输出权值较小,也不足以影响最近的样本,如图中的隐节点3;或者移到了样本有效输入区域外部而不能影响训练样本,如图中的隐节点4.

如果两个隐节点出现了重合,则这两个隐节点对所有样本的响应矢量 $\mathbf{p}_1, \mathbf{p}_2$ 的夹角余弦 $\frac{\mathbf{p}_1^T \mathbf{p}_2}{\|\mathbf{p}_1\| \|\mathbf{p}_2\|}$ 接近于1.图6(b)显示了余下24个被激活的隐节点.图中隐节点1和2对所有样本的输出对应的夹角余弦值为0.9981,而由图可见,这也是互为最近的两个隐节点,它们之间的距离仅为0.0084.另外,这两个隐节点的扩展常数分别为0.1642和0.1625,因此可认为这两个隐节点是重合的.

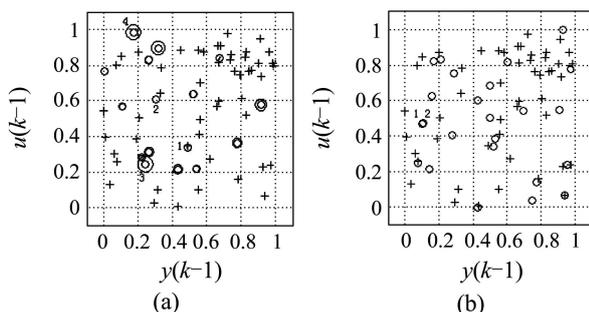


图6 学习后隐节点位置统计

Fig. 6 Statistics of hidden units after learning

为了统计冗余隐节点各种学习动态的出现情况,按上述方式作了100次试验,并统计了100次试验中出现未激活隐节点和重合隐节点的情况.统计结果见表1.试验时取隐节点对样本的激活阈值为0.0001.如果两个激活隐节点对样本输出矢量的夹角余弦大于0.985,且这两个隐节点之间的距离小于0.1,则这两个隐节点为重合隐节点.由表1可见,在RBF网中存在冗余隐节点时,隐节点的萎缩、外移、衰减和重合是频繁出现的现象.

表1 学习动态统计

Table 1 Statistics of learning result

平均未激活隐节点数	17.2100
平均重合隐节点数	2.2800
平均训练误差	0.0751

5 讨论(Discussion)

本文讨论了最小二乘意义下梯度算法收敛后RBF网隐节点参数的取值,揭示了冗余隐节点的萎缩、外移、衰减和重合是频繁出现的现象.这一事实建议笔者在RBF网的训练过程中动态调整网络结构,即根据隐节点位置动态增加,删除或合并隐节点.文献[8]根据这一机理给出了一种实现非线性系统RBF网在线建模的方法.

本文没有分析导致上述现象的原因.最近的分析发现,由于对称性,层次模型(hierarchical model)如RBF网、MLP、高斯混合模型等都属于奇异模型,即这些模型的参数空间存在着奇异点(Singularity),而正是这些奇异点导致了許多奇怪的学习动态,如本文中RBF网隐节点的重合和衰减等^[9].奇异性对层次模型学习动态的影响已成为机器学习和统计学习理论的一个研究热点^[10].

另外,本文所使用的目标函数是最常见的误差平方和.使用其他目标函数(如加上正则化项)是否还会产生类似的学习动态,也是进一步研究的内容.

参考文献(References):

- [1] 魏海坤,徐嗣鑫,宋文忠.神经网络的泛化理论和泛化方法[J].自动化学报,2001,27(6):806-815.
(WEI Haikun, XU Sixin, SONG Wenzhong. Generalization theory and generalization methods for neural networks[J]. *Acta Automatica Sinica*, 2001, 27(6): 806-815.)
- [2] 魏海坤.神经网络结构设计的理论与方法[M].北京:国防工业出版社,2005,2.
(WEI Haikun. *Theory and Methods for Neural Network Design*[M]. Beijing: National Defense Industry Press, 2005, 2.)
- [3] HAYKIN S. *Neural Networks: A Comprehensive Foundation*[M]. New York, NY: Prentice Hall, 1997.
- [4] KARAYIANNIS N, RANDOLPH-GIPS M. On the construction and training of reformulated radial basis function neural networks[J]. *IEEE Trans Neural Networks*, 2003, 14(4): 835-846.

从图6中可看出输出电压的第5、第7次谐波的幅值基本上为零. 由于本算例是以单相系统为控制对象, 选择消除第5、第7次谐波时3次及3的倍数次谐波依然存在. 而实际应用一般考虑三相系统, 当本算例用于三相系统时, 3次及3的倍数次谐波自然消除, 故不需考虑选择消除第3次谐波.

6 结论(Conclusions)

根据对称多项式理论, 将消谐多项式方程组表示成初等对称多项式的多项式方程组大大降低了消谐多项式方程组的次数, 有利于降低消谐多项式方程组的求解难度. 3个开关角的消谐多项式的求解流程和结果表明, 只需计算一些代数表达式就可得到选择性消谐多项式方程组的所有解, 大大减少了计算量, 提高了在线计算的速度.

参考文献 (References):

- [1] PATEL H, HOFT R. Generalized techniques of harmonic elimination and voltage control in thyristor inverters: Part I-harmonic elimination[J]. *IEEE Trans on Industry Applications*, 1973, 9(4): 310 – 317.
- [2] CHIASSON J, TOBERT L M, MCKENZIE K, et al. Eliminating harmonics in a multilevel using resultant theory[C]//*Proc of IEEE Power Electronics Specialits Conf*. Cairns, Australia: [s.n.], 2002: 503 –

508.

- [3] 李师正. 多项式代数[M]. 山东: 山东人民出版社, 1984: 8 – 24. (LI Shizheng. *Polynomial Algebra*[M]. Shandong: Shan Dong People Press, 1984: 8 – 24.)
- [4] 张禾瑞, 郝炳新. 高等代数[M]. 北京: 高等教育出版社, 1984: 90 – 98. (ZHANG Herui, HAO Bingxin. *Advanced Algebra*[M]. Beijing: Advanced Education Press, 1984: 90 – 98.)
- [5] 吴文俊. 几何定理机器证明的基本原理(初等几何部分)[M]. 北京: 科学出版社, 1984. (WU Wenjun. *The Basic Principles of Mechanized Proof of Geometry Theory*[M]. Beijing: Science Press, 1984.)
- [6] 梁宝钰. 用置换法求解二次和三次方程[J]. 太原教育学院学报, 2003, 3: 68 – 70. (LIANG Baoyu. To solve quadratic equation and cubic equation with permutation method[J]. *J of Education Institute of Taiyuan University*, 2003, 3: 68 – 70.)

作者简介:

郑春芳 (1978—), 女, 博士, 研究方向为高频开关电源的谐波消除, E-mail: zcf219@163.com;

张波 (1962—), 男, 教授, 博士生导师, 研究方向为高频开关电源与电力传动, E-mail: epbzhang@scut.edu.cn;

丘东元 (1972—), 女, 副教授, 研究方向为电力电子装置, E-mail: epdyqu@scut.edu.cn.

(上接第360页)

- [5] PLATT J. A resource-allocating network for function interpolation[J]. *Neural Computation*, 1991, 3(2): 213 – 225.
- [6] YINGWEI L, SUNDARARAJAN N, SARATCHANDRAN P. A sequential learning scheme for function approximation and using minimal radial basis neural networks[J]. *Neural Computation*, 1997, 9(2): 1 – 18.
- [7] MICCHELLI C. Interpolation of scattered data: distance matrices and conditionally positive definite functions[J]. *Constructive Approximation*, 1986, 2(1): 11 – 22.
- [8] 魏海坤, 宋文忠, 李奇. 非线性系统RBF网在线建模的资源优化网络方法[J]. 自动化学报, 2005, 31(6): 970 – 977. (WEI Haikun, SONG Wenzhong, LI Qi. Resource-optimizing networks for nonlinear system on-line modeling using radial basis function networks[J]. *Acta Automatica Sinica*, 2005, 31(6): 970 – 977.)

- [9] WEI H, AMARI S. Online learning dynamics of radial basis function neural networks near the singularity[C]//*Proc of Int Joint Conf on Neural Networks*. New York, USA: IEEE Press, 2006: 4770 – 4776.
- [10] AMARI S, PARK H, OZEKI T. Singularities affect dynamics of learning in neuromanifolds[J]. *Neural Computation*, 2006, 18(5): 1007 – 1065.

作者简介:

魏海坤 (1971—), 男, 副教授, 目前在日本Riken脑科学研究所从事博士后研究工作, 最近研究方向为奇异点附近层次神经模型的学习动态, E-mail: hkwei@seu.edu.cn;

李奇 (1963—), 男, 博士生导师, 研究领域为复杂系统过程控制;

宋文忠 (1936—), 男, 教授, 博士生导师, 研究领域为复杂系统辨识及控制、离散事件动态系统.