

基于优化的 RBF 神经网络的变量筛选方法^①

徐富强, 刘相国

(巢湖学院 数学系, 巢湖 238000)

摘要: 由于 RBF 神经网络结构简单、输出与初始权值无关、自适应、可调参数少等特点, 提出了利用交叉验证法寻最优参数 SPREAD 值, 构建最优 RBF 神经网络模型并结合 MIV 算法用于变量筛选。通过实例检验了模型的有效性, 也使该方法具有较好的稳定性和应用性。

关键词: RBF 神经网络; 参数优化; 交叉验证法; MIV; 变量筛选

Variables Screening Methods Based on the Optimization of RBF Neural Network

XU Fu-Qiang, LIU Xiang-Guo

(Department of Mathematics, Chaohu College, Chaohu 238000, China)

Abstract: Because of the Characteristics of RBF neural network structure is simple, output and initialized weights irrelevant, adaptive, less adjustable parameter etc. This paper proposes using the method of cross validation to find the optimal parameter value of SPREAD, constructs the optimal RBF neural network model and combines the algorithm of MIV to use for variables screening. Through the example test the validity of the model, also make the method has better ability of stability and applied.

Key words: RBF neural network; parameter optimization; cross validation method; MIV; variables screening

1 引言

在许多实验条件下, 会产生大量的实验数据^[1-3], 如文献[1,2]对近红外光谱数据分析和文献[3]采用 HPLC 法对甘草样本进行分析所获得图谱数据。而在类似这些数据基础上我们往往需要建立自变量与因变量之间的函数关系, 有时候这种函数关系是非线性的、高度不确定的。当我们对数据进行回归分析时, 就需要建立自变量对因变量的回归模型, 这种回归模型首先需要选择合适的自变量。一般而言, 实验获得的自变量往往较多, 有些自变量对因变量可能影响较小或根本没有影响, 甚至某些变量的获取需要付出较大代价。如果这些变量不重要, 将它们引入模型会降低模型的精度, 达不到理想效果。同时, 大量的自变量间也可能存在多重共线性。所以建模前对自变量的筛选工作十分必要^[4]。由于径向基 (RBF) 神经网络结构简单、训练简洁且学习收敛速度快, 能够逼近任意非线

性函数, 因此本文提出了一种基于优化的 RBF 神经网络的变量筛选方法。

2 径向基神经网络概述

径向基神经网络又称 RBF 神经网络, 属于前向型神经网络。网络的结构与多层前向网络类似, 是一个只有一个隐藏层的三层前向网络。第一层为输入层, 有信号节点组成; 第二层为隐藏层, 该层节点视问题而定, 该层神经元的转化函数即径向基函数是局部响应的高斯函数; 第三层为输出层。RBF 神经网络结构见图 1。

RBF 网络就是用径向基作为隐单元的“基”构成隐藏层, 对输入矢量进行变换, 将低维模式的输入数据变换到高维空间内, 这样就将低维空间内线性不可分的问题转化到高维空间内线性可分, 问题得以解决。

① 基金项目: 巢湖学院自然科学研究资助项目(XLY-201101)

收稿时间: 2011-06-24; 收到修改稿时间: 2011-07-17

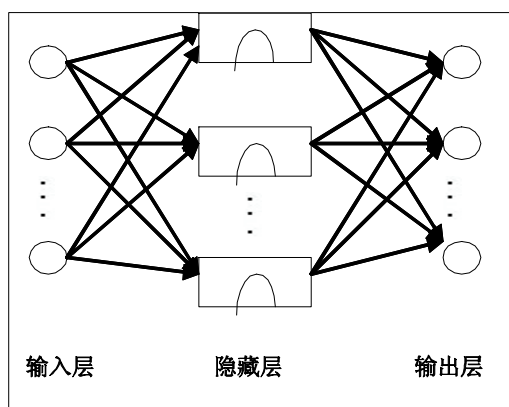


图 1 RBF 神经网络结构

RBF 神经网络学习方法有多种, 本文利用自组织选取中心学习法, 该学习过程分两步: 第一步是自组织学习阶段, 求解隐藏层基函数的中心和方差; 第二步求解隐藏层到输出层间的权值。在训练过程中, 隐藏层神经元数量的确定是一个关键问题, 基本原理是从 0 个神经元开始训练, 通过检查输出误差是否满足条件, 否则自动增加神经元, 每次循环使用, 使网络生成的最大误差对应的输入矢量作为权值向量, 产生一个新的神经元, 然后检查新网络的误差, 重复此过程, 直到误差达到要求或达到最大隐藏层神经元数为止。由此可见, RBF 神经网络具有结构简单、自适应、输出与初始权值无关等特点, 且其学习收敛速度快, 能够逼近任意非线性函数^[5]。

3 参数优化

参数 SPREAD 为径向基函数的扩展速度。在构建好 RBF 神经网络后, 网络学习训练时, SPREAD 的大小直接影响网络的拟合 (预测) 精度, 合理选择 SPREAD 值是很重要的。SPREAD 值越大, 能使径向基神经元对输入矢量所覆盖的空间都产生响应, 但也不需要所有径向基神经元都产生响应, 只要部分响应就足够了, 而且 SPREAD 值太大也会导致计算上的困难。在网络设计时往往都是选取不同的 SPREAD 值进行尝试, 具有一定的主观性和不确定性, 不易得到最优模型^[6,7]。

为了寻求最优模型, 可以将交叉验证方法引入 RBF 神经网络的训练过程, 通过交叉验证, 计算每次网络误差, 循环训练, 找到使网络误差最小对应的 SPREAD 值, 从而找到既定条件下的最佳 SPREAD 值。

4 基于RBF神经网络的变量筛选方法

平均影响值 (MIV) 是评价各个自变量对因变量影响的重要性指标之一。Dombi 等人将 MIV 引入神经网络, 提出了用 MIV 反映网络中权重矩阵的变化情况。目前, MIV 被认为是在神经网络中评价变量相关性最好的指标之一。

将 MIV 引入 RBF 神经网络, 用 MIV 作为评价输入神经元对输出神经元影响大小的一个指标。其符号代表输入对输出相关的方向, 绝对值大小代表相对影响的重要程度。在 RBF 网络训练结束后, 将原训练样本中每一自变量在原值基础上 $\pm 10\%$, 构成两个新的训练样本 K_1 和 K_2 , 将他们分别作为新的训练样本代入已建立的 RBF 网络进行仿真, 分别得到两个输出结果 S_1 和 S_2 , 计算 S_1 和 S_2 的差值, 作为变动后自变量对输出产生的影响变化值 (IV), 最后按训练样本容量计算 IV 的平均得出该自变量对输出变量 (因变量) 的 MIV。按上述方法依次计算每个自变量的 MIV, 根据 MIV 的绝对值大小对各自变量排序, 得到各自变量对因变量影响的相对重要性的次序, 也就实现了变量筛选。

5 应用实例

5.1 数据准备

在区间 $[-2, 2]$ 上分别随机生成 401 个 x_1 、 x_2 和 x_3 值, 并按非线性函数

$$Y = x_1^2 + 2x_2^2 + 5x_3^2 + 2x_1x_2 + 6x_2x_3 + 2x_1x_3$$

产生 Y 值。同时, 在区间 $[-2, 2]$ 上分别随机产生 401 个 x_4 、 x_5 值作为干扰变量。保存 x_1 、 x_2 、 x_3 、 x_4 、 x_5 和 Y 值作为 RBF 网络的训练样本, x_i ($i=1, 2, 3, 4$) 作为网络输入 (自变量), Y 为网络输出 (因变量)。通过 MIV 方法结合优化的 RBF 神经网络筛选出对输出结果有主要影响作用的变量。

5.2 模型建立

先利用训练样本设计一个 RBF 神经网络, 输入层神经元为 5 个, 输出层神经元为 1 个, 网络输出误差设定为 0.01, 利用 newrb 创建一个近似 (approximate) 径向基网络, 在网络学习过程中不断尝试增加隐藏层神经元和个数, 直到网络输出误差满足预先设定值为止, 然后, 通过交叉验证寻找最优参数 SPREAD 值, 找到一个最优的 RBF 神经网络, 再经 MIV 算法实现每个输入对输出的 MIV 值。具体算法流程见图 2。

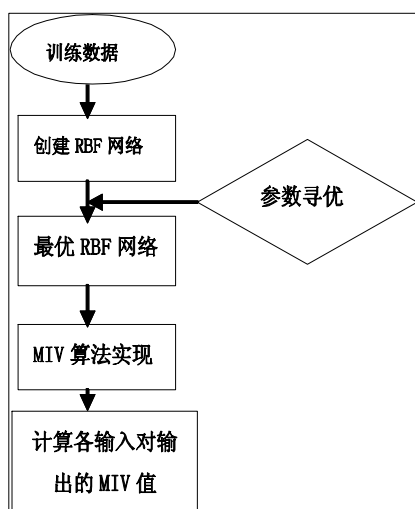


图 2 基于优化的 RBF 神经网络的变量筛选算法流程

5.3 实验结果与分析

参数寻优时, 设定 $SPREAD=0.1:0.1:5$, 重复训练, 经 4 次交叉验证后, $SPREAD$ 值为 3.6 时, 网络均方误差达到最小。将最佳 $SPREAD$ 值代入 RBF 网络中, 建立最优 RBF 神经网络, 当隐藏层神经元达到 356 个时, 网络训练误差达到 0.00995163, 满足误差要求, 学习停止 (见图 3)。再进入 MIV 算法实现, 程序运行结果见图 4。

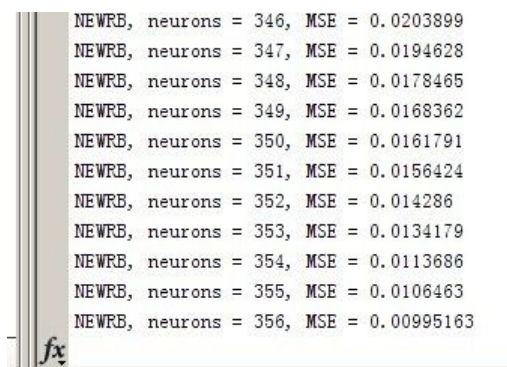


图 3 网络学习过程

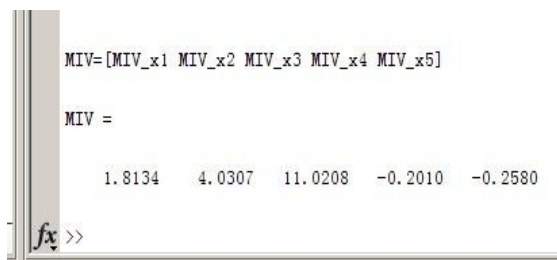


图 4 MIV 算法结果

结果中 MIV_值对应第个自变量对网络输出的 MIV 值, 分别为 1.8134、4.0307、11.0208、-0.2010、-0.2580。符号代表相关的方向, 绝对值大小代表影响的相对重要性。由此可见, 前三个变量、的 MIV 值较大。这是因为值是根据、和计算得出的, 与、无关。所以本文 IMV 算法筛选出对因变量重要影响的结果与真实情况一致。

6 结论

由于 RBF 网络的训练结果与参数 $SPREAD$ 值的选取有较大关系, 本文提出了一种基于优化的 RBF 神经网络的变量筛选方法, 通过交叉验证寻得最优 $SPREAD$ 值, 构建最优 RBF 网络。结合 MIV 变量筛选算法, 获得自变量对因变量的影响重要性评价指标。通过实例分析, 也说明经过优化的 RBF 网络结合 MIV 算法用于变量筛选是可行的。

另外, 相比较 BP、ELman 等其他神经网络, 由于 RBF 网络结构简单、自适应、输出与初始权值无关等特点, 其可调参数少。唯一参数 $SPREAD$ 值经寻优能建立最优 RBF 神经网络, 使模型具有较强的稳定性和可靠性。在对数据进行回归、分类等模型应用方面, 都可以运用该方法进行变量筛选, 建立自变量更少、效果更好的模型。

参考文献

- 褚小立, 袁洪福, 王艳斌, 等. 遗传算法用于偏最小二乘方法建模中的变量筛选. 分析化学研究简报, 2001, 29(4): 437-442.
- 王国庆, 邵学广. 离散小波变换--遗传算法--交互检验法用于近红外光谱数据的高倍压缩与变量筛选. 分析化学研究报告, 2005, 33(2): 191-194.
- 徐恒. 基于统计学的近红外光谱变量筛选与稳健建模方法研究. 天津: 南开大学, 2010.
- 林燕. 偏最小二乘变量筛选法及其应用研究[硕士学位论文]. 厦门: 厦门大学, 2007.
- 葛哲学, 孙志强著. 神经网络理论与 Matlab r2007 实现. 北京: 电子工业出版社, 2007. 117-120.
- 魏海坤, 李奇, 宋文忠. 梯度算法下 RBF 网的参数变化动态. 控制理论与应用, 2007, 24(3): 356-365.
- 彭晓燕, 谭震, 陈昌荣. RBF 网络参数优化方法及其在开关磁阻电机建模中的应用. 科技导报, 2010, 28(19): 42-45.