基于遗传算法的RBF-PLS方法在辐射源识别中的应用

寇 华　　王宝树
(西安电子科技大学计算机学院　西安　710071)

摘　要：RBF-PLS是一种有效的径向基网络构造方法，较好地解决了隐单元数和各中心的取值问题，但宽度系数和PLS成分数难以选定。为此，该文提出采用混合编码遗传算法，以径向基网络的拟合性能和泛化能力为目标，优选宽度系数和PLS成分数，以此建立RBF-PLS-GA模型。将该方法用于雷达辐射源识别，效果良好，明显优于其他网络模型。

关键词：辐射源识别；径向基网络；偏最小二乘回归；遗传算法
中图分类号：TP391.4　　**文献标识码**：A　　**文章编号**：1009-5896(2007)05-1031-04

The RBF-PLS Approach Based on Genetic Algorithm and Its Application in Radar Model Recognition

Kou Hua　　Wang Bao-shu
(*College of Computer Science and Technology*, *Xidian Univ.*, *Xi'an* 710071, *China*)

Abstract: Radial Basis Function-Partial Least Square regression (RBF-PLS) approach is a rapid and efficient method in constructing Radial Basis Function Network (RBFN), and it has put forward a solution to the problem about the choice of the number and the centers of the radial basis functions. But it is difficult to optimize the spread parameter of the radial basis functions and the number of PLS components extracted. A hybrid coding genetic algorithm, which uses different coding methods for different type of variables is proposed to get the optimal solution for the spread parameter and the number of PLS components. The object function of GA is the performance of fitting and predicting of the model. The approach is successfully applied to radar model recognition.

Key words: Radar recognition; Radial basis function network; Partial least squares regression; Genetic algorithm

1　引言

径向基函数网络[1](Radial Basis Function Networks, RBFN)是一种性能优良的前传型网络，被广泛应用于模式识别、非线性系统建模、信号处理和控制等领域。径向基函数(RBF) 网络为三层结构：输入层、隐含层和输出层。输入层接受信源，隐含层对信源作活化处理后传给输出层，经线性处理后即为网络输出。其中，隐层节点的数目、隐层基函数的中心和宽度对网络的性能具有重要的影响。 目前常用的基于聚类的方法，需预先指定类别数 q ，而 q 的选取会影响聚类性能。正交二乘回归方法[2](Orthogonal Least Square Regression, OLSR)用正交化方法筛选回归向量，往往会丢失部分信息而影响网络性能。

本文采用遗传算法[3](Genetic Algorithm, GA)和偏最小二乘回归方法[4,5](Partial Least Square Regression, PLSR)构建径向基网络。将样本容量作为隐单元数，以每个样本个体作为各径基函数的中心，使用遗传算法优选各径基函数的宽度和PLS主成分数 α ，采用偏最小二乘回归方法计算输出层的连接权值，从而避免自变量间的复共线性对模型性能的影响。计算机仿真表明，采用这种混合算法设计的RBF网络具有良好的性能。

2005-09-26 收到，2006-04-24 改回
国家部级基金资助课题

2　径向基函数网络

RBF网络是单隐层前馈网络，如图1所示。

其中 $\boldsymbol{X}=(x_1, x_2,\cdots,x_m)^{\mathrm{T}}\in\boldsymbol{R}^m$ 为输入变量， $\boldsymbol{W}=\{w_{ik}\}_{1\le i\le p,1\le k\le n}$ 为输出层权矩阵，$\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}\in\boldsymbol{R}^n$ 为输出变量。径向基函数常取高斯函数，记为 $\phi_i(\boldsymbol{X})$ ， $i=1,2,\cdots,P$ 。 P 为隐含层径向基函数的个数。第 j 个输入样本在第 i 个隐单元产生的输出为

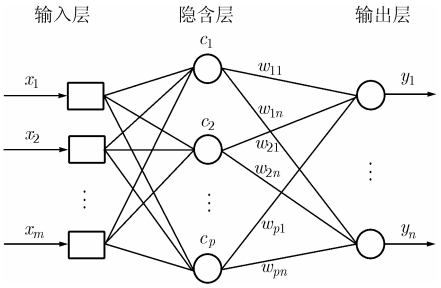

图1　RBFN结构

$$\phi_i\left(\boldsymbol{x}^{(j)}\right)=\frac{\boldsymbol{R}_i(\boldsymbol{x}^{(j)})}{\sum_{q=1}^{P}\boldsymbol{R}_q(\boldsymbol{x}^{(j)})},\quad i=1,2,\cdots,P \tag{1}$$

$$\boldsymbol{R}_i(\boldsymbol{x}^{(j)}) = \exp\left(-\frac{\| \boldsymbol{x}^{(j)} - \boldsymbol{c}_i \|^2}{2\sigma_i^2}\right), \quad i = 1,2,\cdots,P \tag{2}$$

其中 $\boldsymbol{c}_i$ 为第 i 个径向基函数的中心，σ_i 为第 i 个径向基函数的宽度。$||*||$ 多为欧基里德范数。网络第 k 个输出单元的输出为

$$y_k = \sum_{i=1}^{P} w_{ik}\varphi_i(\boldsymbol{X}), \quad k = 1,2,\cdots,n \tag{3}$$

RBFN的性能取决于它的结构和参数，其中输入层结点数 m 和输出层结点数 n 由训练样本数据确定，隐结点数 P，各隐结点径向基函数的参数 $\boldsymbol{c}_i$ 和 σ_i 的取值则由训练算法决定。

3 偏最小二乘回归(Partial Least Square Regression, PLSR)方法

偏最小二乘回归方法是建立在PCA算法之上的应用最为广泛的多元分析方法之一[6]。为了建立由各因素构成的数据矩阵 $\boldsymbol{X}$ 与由各目标构成的数据矩阵 $\boldsymbol{Y}$ 之间的关系，其中 $\boldsymbol{X}$ 包含 p_1 个变量，$\boldsymbol{Y}$包含 p_2 个变量，样本数为 m，传统的处理方法是用最小二乘法建立以下线性模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{4}$$

其中 $\boldsymbol{E}$ 为残差阵，回归系数 $\boldsymbol{B}$ 的最小二乘解为

$$\boldsymbol{B} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{5}$$

用PLS方法处理以上问题时，首先将 $\boldsymbol{X}$ 矩阵作双线性分解，即

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{F} \tag{6}$$

其中矩阵 $\boldsymbol{T}$ 含有两两正交的隐变量或得分矢量 $\boldsymbol{t}$。PLS方法与主成分分析法的不同之处在于，主成分分析法要求分解后得到的隐变量 t 的方差为最大，而不考虑矩阵 $\boldsymbol{Y}$ 的关系。而用PLS方法时，需要用到矩阵 $\boldsymbol{Y}$ 中的信息，矩阵 $\boldsymbol{Y}$ 也可作双线性分解，即

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{E} \tag{7}$$

其中 $\boldsymbol{U}$ 矩阵包含 $\boldsymbol{Y}$ 的隐变量 $\boldsymbol{u}$，即 $\boldsymbol{u}$ 为矩阵 $\boldsymbol{Y}$ 中变量的线性组合，$\boldsymbol{E}$ 为残差阵。PLS方法要求 $\boldsymbol{X}$ 分解得到的隐变量与 $\boldsymbol{Y}$ 分解得到的隐变量相关性为最大。因此有

$$\boldsymbol{u} = \boldsymbol{v}\boldsymbol{t} + \boldsymbol{e} \tag{8}$$

式中 $\boldsymbol{e}$ 为残差矢量，系数 $\boldsymbol{v}$ 根据最小二乘法确定。

Walczak和Massart[5]将PLSR方法引入RBFN的学习，它取隐结点数 P 为训练样本的容量，以每个样本个体作为各径基函数的中心，于是样本容量就等于回归变量数，就会将一些代表噪音的主成分加到模型中，使模型的预测能力下降，从而产生过度拟合(Overfit)现象。为了消除复共线性对RBF网络性能的影响，采用PLSR方法计算权矩阵 $\boldsymbol{w}$。RBF-PLS[5]方法较好地解决隐结点数 p 与各径基函数 $\boldsymbol{c}_i$ 的取值，但它仍用尝试法选定各径基函数的宽度参数，用交叉验证方法确定参与回归的主成分数，计算量较大，且难以选到较优值。本文拟采用遗传算法，优选宽度参数 σ_i 和PLS主成分数 α，以实现高效全局优化。

4 基于遗传算法的RBF-PLS方法的设计

在RBF-PLS方法中，隐结点数 P 取值为训练样本的容量，如果直接对 σ_i 和PLS主成分数 α 编码，用GA 同时搜索 σ_i 和 α 的最优值，则计算量较大。本文采用间接方法，引入参数 N， b， k 和 q，用式(9)，式(10)和式(11)计算 σ_i。

$$\mu_i(\boldsymbol{x}_j) = \frac{\left(1/\left\|\boldsymbol{x}_j - \boldsymbol{x}_i\right\|^2\right)^b}{\sum_{x_c \in \theta_i}\left(1/\left\|\boldsymbol{x}_c - \boldsymbol{x}_i\right\|^2\right)^b}, \quad i = 1,2,\cdots,P \quad \boldsymbol{x}_j \in \theta_i \tag{9}$$

$$\delta_i^2 = \sum_{x_j \in \theta_i} \mu_i(\boldsymbol{x}_j)\left\|\boldsymbol{x}_j - \boldsymbol{x}_i\right\|^2, \quad i = 1,2,\cdots,P \tag{10}$$

$$\sigma_i^2 = k(\delta_i^2)^q, \quad i = 1,2,\cdots,P \tag{11}$$

其中 θ_i 表示与 $\boldsymbol{x}_i$ 最近的 M ($M \le N$)个与 $\boldsymbol{x}_i$ 同类的样本所组成的集合。

集合 θ_i 可由以下方法求出：

(1)令 $j := 1$， $\theta_i := \phi$;设所有训练样本组成的集合为 T。

(2)计算 $\min_{x \in (T-\theta_i-\{\boldsymbol{x}_i\})}\left(\left\|\boldsymbol{x}-\boldsymbol{x}_i\right\|\right)=\left\|\boldsymbol{x}^j-\boldsymbol{x}_i\right\|$， $\boldsymbol{x}^j \in (T-\theta_i-\{\boldsymbol{x}_i\})$

即从不属于 θ_i 且不等于 $\boldsymbol{x}_i$ 的样本中找到距离 $\boldsymbol{x}_i$ 最近的样本 $\boldsymbol{x}^j$。如果 $\boldsymbol{x}^j$ 与 $\boldsymbol{x}_i$ 属于同一类，则将 $\boldsymbol{x}^j$ 加入集合 θ_i 中，转(3)；否则，如果 $j = 1$，生成样本点 $\boldsymbol{x}$，令 $\boldsymbol{x} = \boldsymbol{x}_i + (\boldsymbol{x}^j - \boldsymbol{x}_i)/3$，将 $\boldsymbol{x}$ 加入集合 θ_i 中，停止；如果 $j > 1$，停止。

(3) $j := j + 1$；如果 $j > N$，停止。否则，转(2)。

δ_i^2 可以看作 θ_i 中所有样本到 $\boldsymbol{x}_i$ 的距离的加权平方和。θ_i 中所有样本的权值之和等于1，即

$$\sum_{x_j \in \theta_i} \mu_i(\boldsymbol{x}_j) = 1 \tag{12}$$

σ_i 为第 i 个样本径向基的宽度。b， k， q 为调节 σ_i 大小的可控参数。本文用遗传算法同时搜索 N， b， k， q 和 α 的全局最优值。

4.1 染色体编码

本文以混合方式为染色体基因编码。整型量 N 和 α 的基因采用二进制编码，实型量 b， k， q 的基因采用浮点数编码。

4.2 适应度函数与选择算子

以交叉验证方式建模，即将样本划分为二，其中 k_1 个样本为直接训练样本，其余的 k_2 个样本为检验样本。由种群的每个染色体选定的相关参数建立RBF网络，其中输出层权值待定。输入 k_1 个直接训练样本，得到活化矩阵 $\boldsymbol{A}$。用PLSR方法解回归模型 $\boldsymbol{Y} = \boldsymbol{A} \times \boldsymbol{W}$ 确定输出层权值，PLS主成分数由染色体上 α 值对应的基因经解码给出。计算RBF网络的拟合和预测相对误差：

$$\delta_{\text{self}}=\frac{1}{k_1}\sum_{\boldsymbol{x}_k\in\theta_1}\left(\left\|y^{\text{pre}}(\boldsymbol{x}_k)-y(\boldsymbol{x}_k)\right\|/\left\|y(\boldsymbol{x}_k)\right\|\right) \tag{13}$$

$$\delta_{\text{test}}=\frac{1}{k_2}\sum_{\boldsymbol{x}_k\in\theta_2}\left(\left\|y^{\text{pre}}(\boldsymbol{x}_k)-y(\boldsymbol{x}_k)\right\|/\left\|y(\boldsymbol{x}_k)\right\|\right) \tag{14}$$

其中 θ_1 表示由 k_1 个直接训练样本组成的集合， θ_2 表示由 k_2 个检验样本组成的集合。$y(\boldsymbol{x}_k)$ 表示 $\boldsymbol{x}_k$ 对应的真值，$y^{\text{pre}}(\boldsymbol{x}_k)$ 表示输入 $\boldsymbol{x}_k$ 时，RBF网络对应的输出值。交叉地划分训练样本，使得每个个体都有一次用作检验样本。将每次得到的 δ_{self} 和 δ_{test} 分别作算术平均，记为 $\bar{\delta}_{\text{self}}$ 和 $\bar{\delta}_{\text{test}}$ 。则确定每个染色体的适应度为

$$f=1/(\bar{\delta}_{\text{self}}+m\bar{\delta}_{\text{test}}) \tag{15}$$

m 为权值系数，在训练前根据经验设置，反映了对RBF网络预测能力的重视程度。k_1 和 k_2 的值也需预先制定，如果训练样本比较多，则给 k_2 指定一个较大值，从而减少执行PLSR算法的次数。反之，则给 k_2 指定一个较小值。

本文的选择操作采用基于最佳保留策略的二进制竞争选择(binary tournament)方法和共享函数法[3]。共享函数法可以防止遗传漂移并促进均匀采样。共享函数定义如下：

$$\text{Sh}(d_{ij})=\begin{cases}1-\left(\dfrac{d_{ij}}{\sigma_{\text{share}}}\right)^{a}, & d_{ij}<\sigma_{\text{share}}\\ 0, & \text{其它}\end{cases} \tag{16}$$

其中 d_{ij} 表示当前种群中第 i 个染色体与第 j 个染色体之间的距离。σ_{share} 是小生境半径(niche radius)[3]，要根据所期望的个体之间最小分离程度事先估计出来。一个个体的共享后适应值 f_i' 就由个体适应值 f_i 除以其小生境数 m_i 来确定：

$$f_i'=f_i/m_i \tag{17}$$

给定个体 i 的小生境数 m_i 由在整个种群中对共享函数进行求和得到：

$$m_i=\sum_{j=1}^{\text{pop\_size}}\text{Sh}(d_{ij}) \tag{18}$$

然后用轮盘赌方法[3]相继选出染色体对，从中取出适应值高的染色体放入新种群。如果父代种群中个体的最大适应度大于子代种群中个体的最大适应度，则用父代种群中适应度最大的染色体代替子代种群中适应度最小的染色体。

4.3 交叉算子和变异算子

对于染色体两种编码的基因将分别分段进行交叉和变异操作。对整型基因采用均匀交叉，对实型基因采用改进的算术交叉。算术交叉的公式如下：

子个体1＝父个体1＋ α• (父个体2－父个体1)　　(19)

子个体2＝父个体2＋ α• (父个体1－父个体2)　　(20)

这里， α 是一个比例因子，可由 $[-d,1+d]$ 上均匀分布随机数产生。一般选择 $d=0.25$ 。

在GA中交叉操作被认为是主要搜索算子，因此交叉概率 p_c 一般取较大值，通常取为0.4～0.99。

对整型基因采用基本位变异，对实型基因采用非均匀变异。非均匀变异[3]由Janilow和Michalewiicz提出。它的设计目的是为了得到较高的精确度而具有微调能力。对于给定的父代 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$，如果它的元素 x_k 被选中进行变异，结果的后代 $\boldsymbol{x}'=(x_1,\cdots,x_k',\cdots,x_n)$，其中

$$x_k'=\begin{cases}x_k+\Delta(t,x_k^U-x_k), & \text{random}(0,1)=0\\ x_k-\Delta(t,x_k-x_k^L), & \text{random}(0,1)=1\end{cases} \tag{21}$$

x_k^U 和 x_k^L 分别为 x_k 取值的最大值和最小值。函数 $\Delta(t,y)$ 返回范围 $[0,y]$ 中的一个值，当遗传代数 t 增加时，该值就越来越趋向于0。函数 $\Delta(t,y)$ 的形式如下：

$$\Delta(t,y)=yr(1-t/T)^b \tag{22}$$

其中 r 是[0,1]区间上的随机数， T 是最大遗传代数， b 是确定非均匀程度的参数。变异概率 p_m 一般取较小值，通常取为0.0001～0.5。

5　RBF-PLS-GA在辐射源识别中的应用

根据雷达的特征参数识别出雷达的工作体制是数据融合的研究内容之一。从数据融合系统的知识库中提取80部雷达的特征参数作为样本数据，采用交叉验证方式，每次取出8组数据作为测试样本，其余作为训练样本，共有10次划分。输入矢量为3 维，由雷达的特征参数“重频”、“载频”和“脉宽”组成，输出矢量为8维，分别对应雷达的8种工作体制。

用遗传算法优化RBF-PLS方法时，选定种群规模为40，最大进化代数为100。 交叉操作概率 $p_c=0.7$ ，变异操作概率 $p_m=0.002$ 。图2显示了遗传算法寻找最优解的演化过程。其中 $f(\boldsymbol{x})$ 表示根据种群中适应度最大的染色体构造RBFN的平均误差：

$$f(\boldsymbol{x})=(\bar{\delta}_{\text{self}}+2\bar{\delta}_{\text{test}})/3$$

该图说明当进化到第76代时，找到最优解。

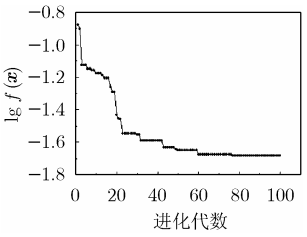

图2　RBF-PLS-GA模型的进化过程曲线

我们分别使用文献[5]介绍的RBF-PLS方法和本文提出的RBF-PLS-GA方法构造径向基网络，选取不同的PLS主成分数，比较它们的拟合误差和预测误差，如图3所示。从中可以看出，RBF-PLS-GA模型的拟合误差和预测误差均小于RBF-PLS模型，这是因为文献[5]提出的RBF-PLS方法要求各径基函数的宽度都相等，极大限制了径基宽度的搜索范围。同时，我们还可以看出，网络的预测误差并不是随着PLS主成分数的增加而单调减小。如果PLS主成分数过多，则会产生过度拟合现象，导致网络的预测能力下降。

表 1 基于 RBF-PLS-GA 方法，RBF-PLS 方法，OLS 方法和 K-Means 的方法的 RBFN 识别效果

		识别率(%)			
		RBF-PLS-GA	RBF-PLS	OLS	K-Means
信噪比	SNR＝20	96.87	96.25	95.31	93.13
	SNR＝12	91.56	89.69	87.81	84.38
	SNR＝10	89.25	87.19	85.47	81.87
	SNR＝8	85.31	82.66	80.63	76.09
	SNR＝5	78.75	75.63	72.19	65.94

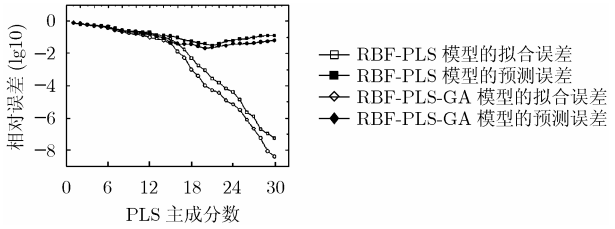

图3 RBF-PLS模型与RBF-PLS-GA模型的拟合误差和预测误差的比较

对训练样本特征参数加上不同程度的高斯白噪声，在不同的噪声环境下，用基于RBF-PLS-GA方法、RBF-PLS方法、OLS方法和K-Means[7]的方法构造RBFN进行识别，其识别结果如表1所示。

由表1可以看出，采用RBF-PLS-GA方法所构造径向基网络的识别性能相对于采用基于RBF-PLS方法，OLS方法和K-Means的方法有了显著提高。用上述80雷达的样本数据训练RBFN时，最终PLS的主成分数只有20，除去了大量噪声，所以该模型的预测性能比较好。

6 结束语

RBF-PLS是一种有效的径向基网络构造方法，但需用尝试法选定宽度系数，而宽度系数的好坏直接影响RBFN模型的性能。本文采用遗传算法，以模型的拟合性能和预报性能为目标，优选宽度系数和PLS主成分数，建立RBF-PLS-GA模型。将这种模型用于雷达辐射源的工作体制识别，效果良好，识别正确率显著高于基于OLS方法和K-Means的构造方法。结果表明，该模型具有良好的拟合性能和泛化能力，这种建模方法对解决其它模式识别问题也有借鉴意义。

参 考 文 献

[1] 吴微. 神经网络计算. 北京: 高等教育出版社, 2003: 41–48.

[2] Chen S, Cowan C F N, and Grant P M. Orthogonal least squares learning algorithm for radial basis function networks[J]. *IEEE Trans. on Neural Networks*, 1991, 2(2): 302–309.

[3] 玄光男, 程润伟. 遗传算法与工程优化. 北京: 清华大学出版社, 2004: 21–108.

[4] 卢涛, 陈德钊. 径向基网络的研究进展和评述. 计算机工程与应用, 2005, 4(19): 60–62.

[5] Walczak B and Massart D L. The radial basis function-partial least squares approach as a flexible non-linear regression technique. *Analytical Chimica Acta*, 1996, 331 (3): 177–185.

[6] Kellner R, Mermet J M, Otto M, and Widmer H M. Analytical Chemistry. New York:Wiley-VCH Verlag GmbH, 1998: 705–727.

[7] Darken C and Moody J. Fast adaptive K-means clustering:some empirical results. Proceedings International Conference on Neural Networks, San Diego, 1990, volume II: 233–238.

寇 华： 男，1980年生，硕士生，研究方向为模式识别、智能信息处理.

王宝树： 男，1940年生，教授，博士生导师，研究方向为信息融合、自动控制、人工智能