基于模糊分组和监督聚类的 RBF 回归性能改进

陈 聪 王士同
(江南大学信息学院 无锡 214122)

摘 要：为了提高 RBF 回归建模的精度，该文提出了一种基于模糊分组和监督聚类的 RBF 回归建模的新方法。基本思想是：首先利用监督聚类将训练样本模糊划分为若干子集，然后分别针对各个子集的样本分布情况进行 RBF 回归建模，最后利用加权组合得到最终的输出。实验表明，该方法对于目标模型的局部细节具有更好的逼近精度。

关键词：径向基函数神经网络；模糊分组；监督聚类；回归
中图分类号：TP183 **文献标识码**：A **文章编号**：1009-5896(2009)05-1157-04

Improved RBF Regression Using Fuzzy Partition and Supervised Fuzzy Clustering

Chen Cong Wang Shi-Tong
(*School of Information Technology*, *Jiangnan University*, *Wuxi* 214122, *China*)

Abstract: In order to improve the precision of RBF regression, this article advances a novel RBF regression modeling method using fuzzy partition and supervised clustering. The proposed method first splits the training data into several subsets using supervised clustering. Then local regression models are independently built with RBF network for each subset. Finally, the output of the network is formed with a weighted combination of each local model. Experiments show that the proposed method achieves more accurate interpretation of local behavior of the target model.

Key words: Radial Basis Function neural Network(RBFN); Fuzzy partition; Supervised clustering; Regression

1 引言

模糊 C 均值法(fuzzy C-means)是一种常用的无监督的聚类方法[1]。无监督的聚类只根据输入数据进行处理，没有考虑输出结果的反馈。目前已提出了许多各具特色的有监督的聚类算法[2−6]，文献[7]提出的利用线性回归模型建立的有监督的聚类，分别考虑了输入、输出空间的情况，具有更高的精度，但是速度较慢。

径向基函数神经网络[8, 9](Radial Basis Function neural Network, RBFN)经常被用在回归问题上。传统的 RBF 回归只使用一个全局模型来描述目标系统，称作“全局建模”，因此对于系统的局部特性逼近能力有限。针对此问题，“局部建模”的方法被提了出来[10]。

本文则把有监督的 RBF 利用文献[10]的思想有机地组合起来，并且在分组过程中也加入监督的成分，根据训练数据的分布情况来确定不同子集的训练方法和训练强度。对于分布简单、容易处理的子集，采用无监督的训练方法，而分布情况复杂的子集采用有监督的训练方法。实验结果表明，此方法既达到了监督聚类的高精度，又提高了处理的速度。

2008-03-31 收到，2008-07-07 改回
国家 863 计划项目(2007AA1Z158)和国家自然科学基金(60773206/F020106)资助课题

2 基于监督聚类的 RBF 回归

在文献[7]中，作者把有监督聚类应用到 RBF 回归问题中，其基本思想是：在 RBF 训练的初始化过程，分别对输入数据和输出数据进行无监督的聚类，求得每个聚类的中心位置。记 $\{x_1,\cdots,x_N\}\subset R_n$ 是 N 个输入数据，$y_k\in R$，$k=1,\cdots,N$ 是相应的输出结果。因此可以建立如下线性回归模型：

$$\hat{y}-z_i=a_i^T(x-v_i),\quad i=1,\cdots,c \tag{1}$$

其中 a_i^T 是第 i 个回归模型的回归参数，v_i 和 z_i 分别是输入空间和输出空间的聚类中心。考虑到数据点相对于聚类中心的隶属度，可以得出如下回归函数：

$$\hat{y}_k=\sum_{i=1}^{c}\hat{y}_k^i=\sum_{i=1}^{c}\mu_{ik}[a_i^T(x_k-v_i)+z_i] \tag{2}$$

这样就可以把输出反馈到聚类过程中，得到新的目标函数：

$$J=\sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{ik}^m\left(\|x_k-v_i\|^2+\alpha\left(y_k-\hat{y}_k\right)^2\right) \tag{3}$$

其中 $\alpha(\alpha>0)$ 是一个影响因子，控制监督项 $\alpha\left(y_k-\hat{y}_k\right)^2$ 对聚类结果的影响程度。

根据式(3)，利用拉格朗日乘数法可以求得监督聚类的迭代公式：

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_k-v_i\|^2+\alpha(y_k-\hat{y}_k)^2}{\|x_k-v_j\|^2+\alpha(y_k-\hat{y}_k)^2}\right)^{1/m-1}},\quad 1\le i\le c,1\le k\le N \tag{4}$$

$$v_i=\frac{\sum_{k=1}^{N}\mu_{ik}^m x_k}{\sum_{k=1}^{N}\mu_{ik}^m},\quad 1\le i\le c \tag{5}$$

3 本文的新方法

传统上使用 RBF 回归建模的一个问题是由于只使用一个模型来描述整个系统的行为，对于系统的局部细节描述能力有限，对于局部特性变化较大的系统往往逼近误差很大。针对此问题，本文借鉴文献[10]的模糊分组的思想，将原训练数据模糊划分成多个子集，利用多个局部模型分别描述整个系统的各个局部行为，从而提高对局部特性的逼近效果。

3.1 算法结构

本文方法的整体结构如图 1 所示。大致过程如下：首先，使用一种新的监督聚类算法对原训练数据进行聚类，从而将其拆分为多个子集。然后对各个子集使用 RBF 进行回归建模，根据各个子集的中训练数据的分布情况来采用有监督或者无监督的训练方法，从而得到多个局部模型。最后将各个局部模型的结果加权组合成最终输出。

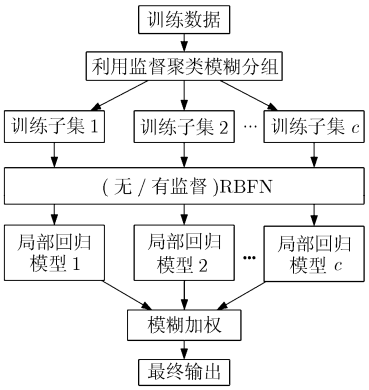

图 1 算法结构

3.2 监督聚类实现模糊分组

本方法的关键问题之一是如何对数据集进行有效的分组。我们希望能够根据各个训练子集回归建模难易程度的不同，分别采用不同的聚类算法进行处理。因此考虑根据训练数据的“波动程度”来进行聚类，把波动程度相似的作为一类，如图 2 的例子。

如果采用无监督的聚类算法(比如 FCM)，把以上训练数据分成两类，因为输入数据完全均匀分布，所以一般会均匀分类。而我们希望把最右边的 4 个点作为一类，因为它们的波动情况相对左边 16 个点来看比较剧烈。因此，改写 FCM 的目标函数：

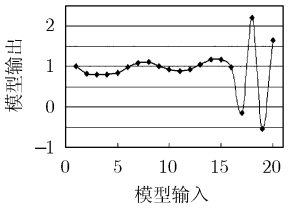

图 2 输入样本的波动情况

$$J=\sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{ik}^m\left(\|x_k-v_i\|^2+\alpha V_i^2\right) \tag{6}$$

其中 $\alpha(\alpha>0)$ 的作用和式(3)中一样是一个影响因子，控制监督项对聚类过程的影响程度。

定义波动指数 V_i，指的是聚类 i 中各个样本点的输出值的“波动情况”：

$$\left.\begin{aligned}V_i&=\sqrt{\sum_{j=1}^{c_i}(D_j-\bar{D}_j)^2\Big/c_i},\quad i=1,\cdots,c\\ D_j&=|y_j-y_{j-1}|,\qquad\qquad j=2,\cdots,c_i\end{aligned}\right\} \tag{7}$$

其中 c_i 指的是第 i 个子集的样本数，D_j 表示第 j 个样本和其相邻样本输出的差值。而 V_i 计算的是第 i 个子集中各个 D_j 的均方差。这样 V_i 就可以表示第 i 个子集输出值的波动情况。

根据式(6)和式(7)，使用拉格朗日乘数法同样可以求出如下迭代公式：

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_k-v_i\|^2+\alpha V_k^2}{\|x_k-v_j\|^2+\alpha V_k^2}\right)^{1/m-1}},\quad 1\le i\le c,1\le k\le N \tag{8}$$

$$v_i=\sum_{k=1}^{N}\mu_{ik}^m x_k\Big/\sum_{k=1}^{N}\mu_{ik}^m,\quad 1\le i\le c \tag{9}$$

经过对原始数据的模糊聚类，可以确定各个聚类的中心，并求得每个聚类的传播宽度：

$$\delta_j^s=\sqrt{\frac{\sum_{i=1}^{p}\mu_{ij}^z\cdot\|x_i^s-v_j^s\|}{\sum_{i=1}^{p}\mu_{ij}^z}},\ s=1,2,\cdots,q,\ j=1,2,\cdots,c,\ q\text{为维数} \tag{10}$$

根据聚类中心和传播宽度，可以判断每个数据点属于哪个子集，如图 3 所示，按照文献[2]的方法，定义一个子集 ψ_k：

$$\psi_k=\left\{(\boldsymbol{x}_i,y_i)\middle|\beta_k^s-\eta\cdot\delta_k^s\le x_k^s\le\beta_k^s+\eta\cdot\delta_k^s,\text{和相应的输出}y_i\right\} \tag{11}$$

其中 $i=1,2,\cdots,N$, $s=1,2,\cdots,q$, $k=1,2,\cdots,c$。η 用来控制子集的叠加程度，η 越大，子集的叠加就越多，子集划分也就越

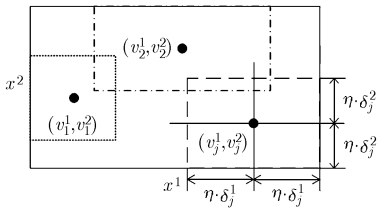

图 3 划分子集

模糊，在接下来的实验中取 3 左右。

接下来针对每个子集分别用一个 RBF 网络进行回归建模，当该子集的波动指数 V_i 大于一定的阈值 V_0 时，选用第 2 节中提到的效果更好的监督聚类算法进行训练，而 V_i 小于 V_0 的子集使用无监督的 FCM 进行训练以求得高效率。这样一来，就可以根据训练样本输出值的分布情况，分段进行处理，兼顾效率与精度。

3.3 加权组合

最终的计算结果由各个局部回归模型(LRM)的结果加权得到：

$$\hat{y}(\boldsymbol{x}_i)=\frac{\sum_{k=1}^{c}W_k(\boldsymbol{x}_i)\mathrm{LRM}^k(\boldsymbol{x}_i)}{\sum_{k=1}^{c}W_k(\boldsymbol{x}_i)},\quad i=1,2,\cdots,N \tag{12}$$

其中第 k 个局部回归模型的权值 W_k 由式(13)得到：

$$W_k(\boldsymbol{x}_i)=w_k^1(x_i^1)\cdot w_k^2(x_i^2)\cdots w_k^1(x_i^q) \tag{13}$$

而每个维度 s 的权值 w_k^s 取决于样本点离聚类中心的距离：

$$w_k^s(x_i^s)=\max\left(\min\left(\frac{x_i^s-(v_k^s-\eta\cdot\delta_k^s)}{v_k^s-(v_k^s-\eta\cdot\delta_k^s)},\frac{(v_k^s+\eta\cdot\delta_k^s)-x_i^s}{(v_k^s+\eta\cdot\delta_k^s)-v_k^s}\right),0\right) \tag{14}$$

其中两个分子 $x_i^s-(v_k^s-\eta\cdot\delta_k^s)$和$(v_k^s+\eta\cdot\delta_k^s)-x_i^s$ 分别计算样本点离子集两个边界的距离，分母是子集的重叠宽度。

4　实验结果及分析

本节将给出一个典型的实验案例以表明本文所提出的方法的有效性。

对波动情况比较复杂的一维函数 $f(x)=e^x\sin(2\pi e^x)/10$ 进行逼近，定义在[−1, 1]，函数曲线如图 4 所示。本实验分别把训练数据划分成 4，6 和 10 个子集，顺序生成 100 对训练数据，精度用归一化的均方差来衡量：$\mathrm{NRMSE}=\sqrt{\sum_{k=1}^{n}\left(Y_k-\hat{Y}_k\right)^2\Big/\left(\sum_{k=1}^{n}\left(Y_k-\overline{Y}\right)^2\right)}$。训练精度的实验结果如表 1 和图 5 所示。

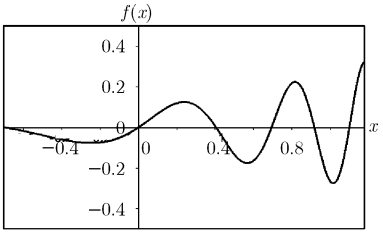

图 4 实验拟合的函数曲线

本实验所逼近的函数波动比较大，因此本方法体现出了明显的优势。可以看出，由于目标系统局部特性波动很大，传统方法在聚类个数比较小的时候很难有效地进行逼近，而本方法很好地描述出了系统的局部特性，训练精度远远好于传统方法，并且子集划分的越多，效果越好。

表 1 训练精度

聚类个数	全局建模	本文方法		
		4 个子集	6 个子集	10 个子集
4	0.96445	0.31891	0.1807	0.03776
5	0.96287	0.24922	0.06943	0.00834
6	0.96291	0.1099	0.00867	0.00434
7	0.81313	0.01948	0.00747	0.00243
8	0.47161	0.01489	0.00999	0.00299
9	0.39659	0.01326	0.01479	0.00279
10	0.35744	0.01314	0.00727	0.00356
11	0.19416	0.01411	0.00843	0.00663
12	0.19462	0.01788	0.01019	0.00478

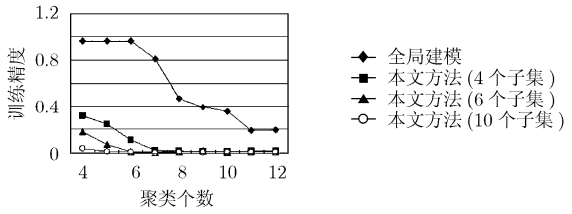

图 5 训练精度

表 2 总结了实验的训练时间，本文方法在速度上和传统方法相当。

表 2 训练时间(s)

聚类个数	全局建模	本文方法		
		4 个子集	6 个子集	10 个子集
4	0.07	0.21	0.271	0.982
5	0.14	0.221	0.37	1.081
6	0.2	0.33	0.391	0.902
7	0.231	0.521	0.35	0.641
8	0.47	0.34	0.351	0.901
9	0.561	0.521	0.39	0.511
10	0.731	0.331	0.381	1.231
11	0.441	0.3	0.391	0.992
12	0.751	0.25	0.45	1.051

5　结束语

本文针对传统 RBF 回归建模方法局部特性逼近精度差的问题，提出了利用模糊分组建立多个局部模型分别逼近目标系统的方法。实验表明，本方法逼近精度远远好于传统方法，尤其是对于输出波动比较复杂的系统。对于本文所给出的方法，尚有下列问题值得进一步研究。

(1)如何确定影响因子 α　有监督聚类中，监督项的影响因子 α 对训练的结果影响很大，其数值需要手工调试，过小，则效果不明显，过大，则容易导致算法不收敛。目前我们采用的方法是手工确定一个初值 α_0，然后按照某一递减函数在每一次迭代中都减少 α 的数值，保证算法最终能够收敛。

(2)如何确定子集的个数　本方法的基本思想是根据训练数据的分布情况来选择子集的个数，把波动情况类似的样本分为一类，如果子集划分得不正确，则最终结果并不理想。而目前如何自动的优化子集个数尚未发现很好的方法。

参 考 文 献

[1] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York: Plenum Press, 1981: 25–78.

[2] Fu Laichung and Wang Shitong. CATSMLP: Toward a robust and interpretable multilayer perceptron with sigmoid activation functions[J]. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 2006, 36(6): 1319–1331.

[3] Pedrycz W. Conditional fuzzy c-means[J]. *Pattern Recognition Lett.*, 1996, 17(6): 625–632.

[4] Pedrycz W. Conditional fuzzy clustering in the design of radial basis function neural networks[J]. *IEEE Trans. on Neural Networks*, 1998, 9(4): 601–612.

[5] Pedrycz W and Vukovich G. Fuzzy clustering with supervision[J]. *Pattern Recognition*, 2004, 37(7): 1339–1349.

[6] Cheng Chibin. Fuzzy regression with radial basis function network[J]. *Fuzzy Sets and Systems*, 2001, 119(2): 291–301.

[7] Antonino Staiano. Improving RBF networks performance in regression tasks by means of a supervised fuzzy clustering[J]. *Neurocomputing*, 2006, 69(13-15): 1570–1581.

[8] 缑水平, 焦李成, 田小林. 基于免疫克隆聚类协同神经网络的图像识别[J]. 电子与信息学报, 2008, 30(2): 263–266.
Gou Shui-ping, Jiao Li-cheng, and Tian Xiao-lin. Image recognition using synergetic neural networks based on immune clonal clustering[J]. *Journal of Electronics & Information*, 2008, 30(2): 263–266.

[9] 唐洪荣, 沈民奋, 李斌. 周期紧支撑径向基函数对 BEMD 的优化[J]. 电子与信息学报, 2008, 30(1): 149–153.
Tang Hong-rong, Shen Min-fen, and Li Bin. The improvement of the BEMD using compactly supported RBF[J]. *Journal of Electronics & Information*, 2008, 30(1): 149–153.

[10] Chuang Chenchia. Fuzzy weighted support vector regression with a fuzzy partition[J]. *IEEE Trans. on Systems, Man, And Cybernetics*, 2007, 37(3): 630–639.

陈　聪：　男，1981 年生，硕士生，研究领域为人工智能、模式识别.

王士同：　男，1964 年生，博士，教授，博士生导师，主要研究领域为人工智能、模式识别、生物信息学.