基于聚类算法的 RBF 神经网络设计综述

张 彬
(长沙理工大学 电气与信息工程学院，湖南 长沙 410004)

摘 要：简要分析了径向基函数(RBF)神经网络。在此基础上，介绍了 K-均值聚类算法的神经网络、C-均值聚类算法的神经网络和 PAM 聚类算法的神经网络三种聚类算法的 RBF 神经网络。展望了基于聚类的 RBF 神经网络设计的发展趋势。

关键词：聚类；RBF 神经网络；发展趋势
中图分类号：TP18 **文献标识码**：A **文章编号**：1674-7720(2012)12-0001-03

Overview on design of RBF network based on fuzzy clustering

Zhang Bin
(School of Electrical & Information Engineering, Changsha University of Science & Technology, Changsha 410004, China)

Abstract: This paper briefly analyzed the mathematical model of RBF neural network. On this basis, three kinds of RBF network based on fuzzy clustering were introduced: K-means clustering algorithm, C-means clustering algorithm and PAM clustering algorithm. Finally, it was expected about the future trend of design of RBF Network based on fuzzy clustering.

Key words: fuzzy clustering; RBF network; the future trends

径向基函数(RBF)神经网络是前向型神经网络[1]，能够以任意精度逼近于任意函数。因为 RBF 网络结构相对简单、非线性逼近能力强和收敛速度较快，现在已经广泛应用在工业智能控制和系统优化、通信系统的信号以及信息处理等诸多领域。对该网络的深入研究也越来越受到国内外学者的共同关注。而聚类分析[2-3]是一种对数据进行分析和建模的重要方法，即将没有明显规律的数据源，依据某些特性，将数据划分到有区别的数据类中，所采用的聚类算法就是聚类分析研究的重点。本文综述了三种聚类算法是如何构造 RBF 神经网络。

1 RBF 神经网络结构及其原理

RBF 神经网络的原理[4]是模拟人脑中局部协调和相互覆盖接收范围的神经网络构造。它是一种三层前向网络，由输入量到输出量的映射存在非线性，而隐含层空间到输出空间的映射却是线性的，从而提高了学习速度，同时也避免了局部极小问题。RBF 神经网络结构如图 1 所示。它具有 n 个输入节点、m 个隐含节点和 1 个输出节点。

网络结构中，假设 $\boldsymbol{x}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}$ 为该网络的输入向量，而设 RBF 网络的径向基向量 $\boldsymbol{H}=[h_1,h_2,\cdots,h_m]^{\mathrm{T}}$，其中 h_j 为高斯基函数：

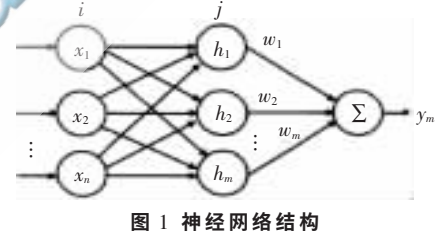

图 1 神经网络结构

$$h_j=\exp\left[\frac{\|\boldsymbol{x}-\boldsymbol{C}_j\|^2}{2b_j^2}\right] \quad (j=1,2,\cdots,m) \tag{1}$$

式(1)中，$\boldsymbol{C}_j$ 为网络的第 j 个节点的中心矢量，$\boldsymbol{C}_j=[c_{j1},c_{j2},\cdots,c_{jn}]^{\mathrm{T}}$。

设 $\boldsymbol{B}$ 为网络的基宽向量：

$$\boldsymbol{B}=[b_1,b_2,\cdots,b_m]^{\mathrm{T}} \tag{2}$$

b_j 为节点 j 的基宽度参数，其值为大于零的数。

设 $\boldsymbol{W}$ 为网络的权向量：

$$\boldsymbol{W}=[w_1,w_2,\cdots,w_j,w_m]^{\mathrm{T}} \tag{3}$$

则网络的输出表示式为：

$$y_m(k)=w_1h_1+w_2h_2+,\cdots,+w_mh_m \tag{4}$$

构造和训练 RBF 神经网络就是要使它经过学习来确定每个隐层神经元基函数的中心和宽度，然后再利用最小二乘或其他方法求出隐含层到输出层的权值向量，从而构建出所研究系统的输入到输出的映射关系。

2 聚类的 RBF 神经网络设计

对 RBF 神经网络来说,隐层节点中心和基函数宽度的初始值的确定会直接影响神经网络的收敛速度,因此选择合适的两个参数的初始值可以提高收敛速度,其选择方法有很多,比如有梯度下降法、模糊理论算法和自适应模糊等方法。而本文就是综述 K-均值聚类算法、C-均值聚类算法和 PAM 聚类算法,利用不同的聚类分析算法得到隐层节点中心和基函数宽度,从而构造和训练出合适的 RBF 神经网络。

2.1 K-均值聚类算法的 RBF 神经网络设计

2.1.1 K-均值聚类算法基本思想[5]

K-means 算法是一种基于误差平方和准则的聚类算法,首先在样本数据中选取 k 个对象作为初始的聚类中心,确定下来后每个点都被分配到与其最近的聚类中心,而到同一个聚类中心的点集被指定为一个分组,即形成 k 个分组;重复分配和更新步骤,直到分组不发生任何变化,也就是聚类中心不发生变化时为止,最后具有共同特性的样本数据就形成了特定的组群。

若 N_i 是第 i 个聚类 Γ_i 中的样本数据,m_i 是该样本的均值,其表达式为:

$$m_i=\frac{1}{N_i}\sum_{y\in\Gamma_i} y \tag{5}$$

将 Γ_i 中的样本 y 与均值 m_i 间的误差平方和对所有类相加,其表达式为:

$$J_e=\sum_{i=1}^{k}\sum_{y\in\Gamma_i}\| y-m_i\|^2 \tag{6}$$

J_e 是误差平方值和数据样本的聚类准则,它是数据样本集 Y 和类别集 Ω 的函数关系式。J_e 度量了用 k 个聚类中心 $m_1,m_2,\cdots,m_k$ 代表 k 个样本子集 $\Gamma_1,\Gamma_2,\cdots,\Gamma_k$ 时刻所产生的总的误差平方。对于不同的聚类,J_e 的值也不同,使 J_e 能够达到极小的聚类就是误差平方和准则下的最优结果。

从它的原理上来讲,该算法可靠、计算简单、收敛速度较快,能比较好地处理较大的数据集,但 K-means 算法对初始聚类中心很敏感,所以从不同的初始聚类中心出发,得到的聚类结果也不一样。

2.1.2 K-means 算法的 RBF 神经网络设计步骤

设 k 为网络迭代次数,第 k 次迭代时聚类中心设为 $c_1(k),c_2(k),\cdots,c_h(k)$,与之相对应的聚类域为 $w_1(k),w_2(k),\cdots,w_h(k)$。K-means 算法确定 RBF 神经网络中心隐层节点中心 c 和基函数宽度 b:

(1) 初始化设定:先选择 h 个不同的初始聚类中心,令 $k=1$。从样本中选择前 h 个样本输入,这 h 个数据中心取值不能相同。

(2) 计算所选取的样本输入与聚类中心的距离,即 $\| X_j-c_i(k)\|$,$i=1,2,\cdots,h$,$j=1,2,\cdots,N$。

(3) 对输入样本 X_j,按距离最小规则对样本进行分类:即当 $i=\min_i \| X_i-c_i(k)\|$,$i=1,2,\cdots,h$ 时,X_j 即被归为第 i 类,即 $X_i\in w_i(k)$。

(4) 重新计算各类新的聚类中心:$c_i(k+1)=\frac{1}{N}\sum_{x\in w_i(k)} x$,$i=1,2,\cdots,h$。

(5) 如果 $c_i(k+1)\neq c_i(k)$,返回到步骤(2),否则聚类结束,转到步骤(6)。

(6) 根据各聚类中心之间的距离确定各隐节点的基宽向量,即 $b_i=\sigma d_i$,其中 d_i 为第 i 个聚类中心与其他最近样本数据中心之间的距离,即 $d_i=\min_i \| c_j-c_i(k)\|$,$\sigma$ 为重叠系数,采用式(1)高斯基函数计算隐藏层节点的输出量。

2.2 C-均值聚类算法的 RBF 神经网络

2.2.1 C-均值聚类算法(FCM)基本思想[6]

FCM 聚类采用最小化误差平方和目标函数的方法建立最优分类。把聚类问题归结成一个约束非线性规划的问题,即将原始问题转化为优化问题,利用非线性规划理论求解获得数据集的模糊划分从而得到数据的聚集。首先,数据样本随机选取聚类中心的数目,并给定所有数据点与聚类中心的模糊隶属度;然后以极小化所有数据点到各个聚类中心的距离与隶属度的加权和为优化目标,通过迭代反复修改聚类中心和分类矩阵,直到满足停止条件。

假设数据样本的集合为 $X=\{x_1,x_2,\cdots,x_n\}$,将数据集合分成 c 个模糊组,求出各组的聚类中心 $c_j(j=1,2,\cdots,c)$,使得目标函数能达到最小。

目标函数表达式为:

$$J_e=\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{\alpha}\| x_i-c_j\|^2,\ 1\leqslant\alpha\leqslant\infty \tag{7}$$

同时需要满足条件式:

$$\sum_{j=1}^{c}\mu_{ij}=1,\ \forall i=1,2,\cdots,n \tag{8}$$

其中,$\mu_{ij}\in[0,1]$ 表示为第 i 个数据点关于第 j 个聚类中心的隶属度;c_j 是第 j 个聚类中心,初始值就在训练数据样本集中随机抽取;α 为模糊度,α 越大聚类的模糊性越大,本文设定 $\alpha=2$。模糊聚类通过多次迭代最优化目标函数 J_e 实现。其中模糊隶属度 μ_{ij} 和聚类中心 c_j 表达式分别为:

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\| x_i-c_j\|^2}{\| x_i-c_k\|^2}\right)^{\frac{2}{(\alpha-1)}}} \tag{9}$$

$$c_j=\frac{\sum_{i=1}^{n}\mu_{ij}^{\alpha}x_i}{\sum_{i=1}^{n}\mu_{ij}^{\alpha}} \tag{10}$$

计算过程从一个随机的聚类中心开始,通过寻找目标函数的最小点,反复调整聚类中心和各个样本的隶属

度，在 J_c 的局部最小点处收敛，最终达到确定样本类别。

该算法也存在着某些不足的地方，如：算法性能取决于初始聚类中心的选取，其聚类的个数是不能自行确定的；它对数据样本中孤立点和噪音数据比较敏感；目标函数的设定没有全面地考虑到数据样本分布不均衡。

2.2.2 C-均值聚类算法的 RBF 神经网络设计步骤

设 k 为迭代次数，则第 k 次迭代时的聚类中心为 $c_1(k),c_2(k),\cdots,c_h(k)$，相对应的聚类域为 $w_1(k),w_2(k),\cdots,w_h(k)$。FCM 算法确定 RBP 神经网络中心 c 和基宽向量 b 的过程为：

（1）选择数据样本的聚类个数 c 和设定其初始化聚类中心，$t=0$；

（2）由式（9）计算得 x_i 在第 j 个聚类中的隶属度；

（3）由式（10）更新聚类中心 c_j，于此同时原来的模糊隶属度 μ_{ij} 也更新为 $\mu_{ij}^{(t)}=t+1$；

（4）由式（7）计算得目标函数 J_c，若结果相对于上次目标函数值的变化量小于某个阈值，则算法结束，否则返回到步骤（3）继续更新神经网络中心 c 和基宽向量 b；

（5）根据各个聚类中心之间的距离大小来确定每个隐节点的基宽向量 $b_i=\sigma d_i$，其中 d 为第 i 个聚类中心与其他最近的数据中心之间的距离，即 $d_i=\min_i \| c_j-c_i(k) \|$，$\sigma$ 为重叠系数。该算法中有两个参数要预先设定：即阈值 ε 和聚类个数 c。阈值 ε 可以人为指定一个较小值，只要能满足所要求的精度即可。聚类个数 c 通常根据经验知识来确定其取值范围。然后由式（1）高斯基函数计算出隐藏层节点的输出。

2.3 PAM 聚类算法的 RBF 神经网络

2.3.1 PAM 聚类算法[7]

基本思想为：选用组中位置最中心的数据点，将 n 个数据点划分为 k 个；代表数据点称为中心点，而其他数据点被称为非代表数据点；最初任意选择 k 个数据点作为中心点，该算法反复地用非代表数据点来代替代表数据点，试着找出更好的中心点，从而改进聚类的质量；在每次迭代过程中，分析所有可能的数据点对，每个对中的一个数据点是中心点，而另一个是非代表数据点。对可能的各种组合，估计聚类结果的质量；一个数据点能被使最大平方-误差值减少的对象所代替；在一次迭代中产生的最优数据点集合就成为下次迭代的中心点。为了找出 k 个中心点，算法先任意地从数据样本中选择 k 个对象。然后用一个非选中数据点 P_h 替换一个选中数据点 P_i，若这种替换能够提高聚类效果，由数据点和它所属组的中心点之间的平均相异度来度量，一般计算欧式距离。

为了得到 P_h 与 P_i 之间替换的效果，算法通过计算每一个非选中数据点 P_j 的代价，即 C_{jih}。就 P_j 的情况，计算 C_{jih} 的表达式如下：

（1）当前 $P_j \in P_i$ 所代表的组，并且 P_j 离 P_{j2} 比 P_h 较近，即 $d(P_j,P_h) \geqslant d(P_j,P_{j2})$，则 P_{j2} 是 P_j 的第二最近中心点。此时若将 P_h 替换 P_i 作为中心点，则 $P_j \in P_{j2}$ 所代表的组，因此就 P_j 来说替换的代价为 $C_{jih}=d(P_j,P_{j2})-d(P_j,P_i)$；

（2）当前 $P_j \in P_i$ 所代表的组，并且 P_j 离 P_h 比 P_{j2} 较近，即 $d(P_j,P_h)<d(P_j,P_{j2})$，则 P_{j2} 是 P_j 的第二最近中心点。此时若将 P_h 替换 P_i 作为中心点，则 $P_j \in P_h$ 所代表的组，替换代价为 $C_{jih}=d(P_j,P_h)-d(P_j,P_i)$；

（3）当前 $P_j \in P_{j2}$ 所代表的组，并且 P_j 离 P_{j2} 比 P_h 较近。此时若将 P_h 替换 P_i 作为中心点，则 P_j 仍然属于 P_{j2} 所代表的组，替换代价为 $C_{jih}=0$；

（4）当前 $P_j \in P_{j2}$ 所代表的组，并且 P_j 离 P_h 比 P_{j2} 近。此时若将 P_h 替换 P_i 作为中心点，则 $P_j \in P_h$，替换代价为 $C_{jih}=d(P_j,P_h)-d(P_j,P_{j2})$。

考虑上述 4 种情况，对所有 $n-k$ 个 P_j 的代价 C_{jih} 求和，用 P_h 替换 P_i 的总代价即为 $TC_{ih}=\sum C_{jih}$，$j=1,2,\cdots,n-k$。

2.3.2 PAM 聚类算法的 RBF 神经网络设计步骤

首先根据高斯径向基函数的特点，即只有小部分靠近中心 c 的输入被激活；考虑分类的特性，即靠近各类中心点的数据点被归为一类。用 PAM 算法对数据样本进行聚类，从聚类结果得到数据样本的各个中心点；然后把各聚类中心作为各径向基函数的 c 值，则靠近 c 的数据点会被激活，而远离 c 的数据点不会被激活，用这种方法能提高分类的精确度；然后得到基函数的宽度，第二阶段利用数据样本求出隐藏层与输出层神经元之间的连接权向量 w，进而完成整个网络的训练过程。

基于 PAM 聚类的 RBF 神经网络训练过程算法具体步骤为：

（1）对数据样本集进行归一化处理，计算数据集中对象两两之间的距离；

（2）在 n 个对象中任意选取 k 个当成初始中心点；

（3）将其他 $n-k$ 个对象划入离其最近的中心点所代表的组中去；

（4）
```
while TRUE do
    minTC=10000;
    for i=1 to k do
        for h=2 to n-k+1 do
        TC=0;
            for j=2 to n-k+1 do
            TC=TC+Cjin;
                if TC<minTC  minTC=TC;
```

使用替换重新形成新的 k 个中心点集合返回到步（3）继续；直到 $\min TC>0$ 时算法停止；

（5）计算各中心点之间的距离，根据式 $b=d_{\max}/\sqrt{2k}$ 计算径向基函数的宽度。采用式（1）高斯基函数计算隐藏层节点的输出。

3 发展趋势

要进一步提高聚类算法对 RBF 神经网络隐层节点

中心和基函数宽度的确定，需要优化改进现有的聚类算法，以提高学习性能。因此，出现了一些在原聚类算法基础上改进的聚类算法[8-10]，弥补了样本分析在聚类过程中存在的某些不足，将聚类算法的性能发挥得更加充分，能更有效地与 RBF 神经网络结合起来。

参考文献

[1] 苏美娟.径向基函数神经网络学习算法研究[D].苏州:苏州大学，2007.

[2] CHIU S L. Fuzzy model identification based on cluster estimation[J]. Journal of Intelligent and Fuzzy System，1994，2(3)：1240-1245.

[3] 满春涛，李晓霞，张礼勇.一种基于 ACO 的 RBF 神经网络训练方法[J].哈尔滨理工大学学报，2008，13(1):59-61,65.

[4] 刘金琨.先进 PID 控制及其 MATLAB 仿真[M].北京:电子工业出版社，2003.

[5] 吴晓蓉.K-均值聚类算法初始中心选取相关问题的研究[D].湖南:湖南大学，2008.

[6] 何迎生，段明秀.基于模糊聚类的 RBF 分类器的设计与实现[J].重庆科技学院报，2009，12(2):101-103.

[7] 段明秀，孙可.基于 PAM 聚类方法的 RBF 神经网络设计[J].沈阳师范大学学报，2009，27(4):440-443.

[8] 朱长江，张缨.模糊 C-均值聚类算法的改进研究[J]. 河南大学学报，2012，42(1):92-95.

[9] 李春富，郑小青，葛铭.基于改进聚类算法的 RBF 网络及其应用[J].南京工业大学学报，2011，33(6):72-76.

[10] 庞振.徐蔚鸿.一种基于改进 K-means 的 RBF 神经网络学习方法[J].(2011-07-14).[2012-01-20]:http://www.cnki.net/kcms/detail/11.2127.TP.20110714.1550.036.html.

(收稿日期:2012-02-27)

作者简介：

张彬，男，1982 年生，硕士研究生，助理工程师，主要研究方向：智能信息处理和智能控制。