浙江理工大学学报(自然科学版),第 31 卷,第 1 期,2014 年 1 月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 31, No. 1, Jan. 2014

文章编号: 1673-3851 (2014) 01-0071-04

基于 OCA 客观聚类的 RBF 神经网络的研究

黄　静, 张　江

(浙江理工大学信息学院, 杭州 310018)

摘　要: 影响 RBF 神经网络性能的关键因素是基函数中心的选取,而目前尚没有可靠的方法选取 RBF 神经网络的中心。基于 GMDH 理论的 OCA 客观聚类具有能够自动确定最优聚类个数的优点。将 OCA 聚类应用于 RBF 神经网络中,用以自适应确定隐节点数目和各径向基函数中心,克服了传统 RBF 网络不能客观确定隐节点数目的缺点。实验仿真结果表明,基于 OCA 客观聚类的 RBF 神经网络具有自适应性、正确率高和训练速度快的优点。

关键词: RBF 神经网络; OCA 客观聚类; 隐节点数目; 基函数中心

中图分类号: TP183　　**文献标志码**: A

0　引　言

RBF 神经网络首先由 Broomhead 和 Lowe 提出。由于其具有结构简单、学习速度快、能够逼近任意非线性函数的优点,因此广泛应用于模式识别、非线性函数逼近等有实时性要求的领域[1]。已有研究表明,影响 RBF 网络性能的关键因素是基函数中心的选取,而一般 RBF 网络所利用的非线性激活函数形式对网络性能的影响并非至关重要。目前,确定 RBF 网络中心的方法主要有:随机选择法、自组织学习法和正交最小二乘法(OLS)等[1-2]。随机选择法选择中心,可能会出现两个中心非常接近的情况,解线性方程组时出现奇异矩阵,导致解的结果不可靠。因此,随机选择法只适用于给定样本数据具有代表性的问题。通过自组织学习进行聚类,选取聚类中心作为 RBF 中心,常用的方法有 K-均值聚类、最近邻聚类和模糊聚类等,但都必须先给出聚类个数。正交最小二乘(OLS)方法源于线性回归模型,是一种应用较多的方法,但是不能进行迭代训练。

乌克兰科学院 Ivakhnenko A G 提出了基于自组织数据挖掘思想 GMDH(数据分组处理方法)的 OCA 客观聚类[3]。OCA 客观聚类能够自适应地确定聚类个数,RBF 通过取各类样本的平均值作为相应隐节点的数据中心。因此,基于 GMDH 的 RBF 神经网络不需要给出聚类个数,并且能够进行迭代训练。本文通过对膨胀土分类实验的仿真,验证了基于 OCA 客观聚类的 RBF 神经网络性能。

1　RBF 神经网络

RBF 神经网络是一种常用的前馈网络,拥有很强的非线性拟合能力,可以映射任意复杂的非线性关系,而且学习速度快,结构简单。Poggio 和 Girosi 已经证明,RBF 网络是连续函数的最佳逼近。RBF 网络采用局部激励函数,很大程度上克服了 BP 神经网络训练过程很长、容易陷入局部极小值的缺点[4]。RBF 神经网络包括三层:输入层、隐层和输出层[5]。输入层由一些感知单元组成,它们将网络与外部环境连接起来。RBF 网络仅有一个隐层,它执行从输入空间到隐藏空间之间进行非线性变换。输出层是线性的,为作用于输入层的激活信号提供响应。

如图 1 所示,输入层有 N 个节点,输入层节点个数等于样本维数。隐层有 P 个节点,各隐节点的基函数的形式为:

收稿日期: 2013－09－12

基金项目: 浙江省自然科学基金资助项目(LY12F03012)

作者简介: 黄　静(1965－),女,杭州人,教授,博士,主要从事图像处理方面的研究。

通信作者: 张　江,E-mail:462098626@qq.com

$$\varphi=(\|X-X^p\|),p=1,2,\cdots,P$$

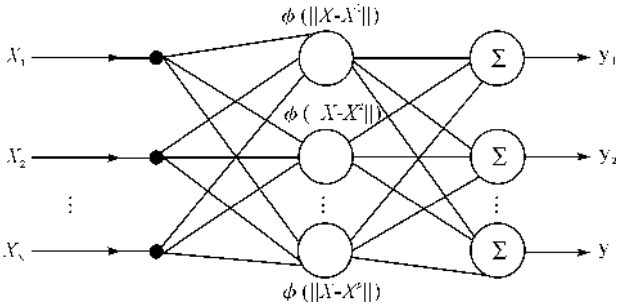

图1 RBF神经网络结构

基函数 φ 为非线性函数，训练数据点 X^p 是 φ 的中心。基函数以输入空间的点 X 与中心 X^p 的距离作为函数的自变量，一般选用 Gauss 函数：

$$\varphi(r)=\exp\left(-\frac{r^2}{2\sigma^2}\right)$$

$Y=(y_1,y_2,\cdots,y_l)^T$ 为网络输出，采用线性激活函数。当输入训练集中地某个样本 X^p 时，对应的期望 d^p 就是教师信号。将训练集样本逐一输入，从而确定网络隐层到输出层之间的 P 个权值。

2 OCA 客观聚类

基于 GMDH 理论的 OCA 客观聚类具有能够自动确定聚类个数的优点。已有研究表明，OCA 能对模糊对象给出更精确地近似或随机过程的预测给出更好的结果[6]。OCA 算法首先计算样本两两之间的距离，构成偶极子。将偶极子分成包含相同样本数的两个子集 A 和 B。同样的方式在剩余样本中，得到子集 C 和 D，作为检测集。然后，对子集 A 和 B 进行聚类，利用一致性准则得到最优方案。如果最优方案多于一个，利用检测集 C 和 D 确定唯一的最优方案[7-10]。假设有 n 个样本 $\{x_1,x_2,\cdots,x_n\}$，样本维数为 m，OCA 聚类具体步骤如下：

a) 构造子集 A、B、C 和 D。

距离计算公式为：$d_{ij}=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$，其中 $i,j\in(1,2,\cdots n)$ 且 $i\neq j$。总共有 C_n^2 个样本距离值，将其按从小到大排列。按一定规则选取$[n/2]$(取整)个距离值 d_{ij}，构成 $p=[n/2]$ 个偶极子，每个偶极子记作(i,j)。其中，选取规则为：从样本距离值的最前端开始，当前要选取的偶极子(i,j) 中的 i 值和 j 值，如果前面都没有出现过，就选取它，否则弃之，即选取 p 个偶极子后的 i 值和 j 值恰好构成整个样本集序号$(1,2,\cdots n)$。按先后顺序取这 p 个偶极子的 i 值，构成子集 A；取其 j 值，构成子集 B。

然后，从剩余的 C_n^2-p 个距离值中，按同样的规则选取 p 个偶极子，分别按先后顺序取其 i 值和 j 值，构成子集 C 和 D 。

b) 对子集 A 和 B 进行聚类。

首先，按先后顺序对子集 A 和 B 中样本点进行编号，如表1所示，每对偶极子所对应的子集 A 和 B 中两个样本点编号相同，即编号都为$(1,2,\cdots,p)$。把子集 A 和 B 中对应的每列称为一类，此时子集 A 和 B 中有 $k=p$ 类。

表1 子集 A 和 B 样本点编号

A	1	2	……	k
B	1	2	……	k

然后，把子集 A 和 B 聚成 $k-1$ 类。具体方法如下：计算子集 A 中任意两个编号样本之间的距离，将距离最小的两个编号归为一类(这两列将拥有两个相同的编号)，假设编号1和编号2距离最小，如表2所示。

表2 编号1,2归为一类

A	1,2	1,2	……	k

同样的方法，将子集 B 聚成 $k-1$ 类。将子集 A 和 B 中对应每列进行比较，相同列的数目记为 Δp。使用一致性准则：$\varepsilon_1=(p-\Delta p)/p$，求出 ε_1 的值。特别指出，若类中不止一个样本点，其它类中任意一个样本点只要和该类中其中一个样本点的距离最小，便认为这两个类距离最小。重复上述步骤，把子集 A 和 B 聚成 $k-2$ 类，$k-3$ 类，……，2类，计算相应的 ε_i 值。

统计 $k-1,k-2,\cdots,2$ 类中 $\varepsilon=0$ 的类数目，记为 $count$，它们成为最优聚类方案的候选方案。显然当 $count=1$ 时，唯一的候选方案即为最优聚类方案。当 $count>1$ 时，就要用到检测集 C 和 D 了。

c) 对子检测集 C 和 D 进行聚类。

聚类过程同步骤 b)一样。考察步骤 b)的几种候选方案，看哪种方案的 ε 在检测集 C 和 D 上最小。这个唯一的聚类方案，便是要找的最优聚类方案。

从上述过程可以看出，OCA 客观聚类能够客观地确定聚类个数，得出最优方案，而不需要有先验知识。

3 基于 OCA 客观聚类的 RBF 神经网络

由于 OCA 聚类能够客观正确地对样本进行聚类，从而获取最佳基函数数据中心。采用 OCA 聚类的 RBF 神经网络实现步骤如下：

a) 确定 RBF 神经网络的训练样本、测试样本、输出层节点数目。样本的维数作为输入层节点数目；

b) 用 OCA 聚类方法对训练样本进行聚类,得出分类个数,作为隐节点数目;

c) 分别计算每类的数据中心,即平均值,作为相应隐节点的基函数数据中心;

d) 利用训练样本训练隐层到输出层的权值;

e) 利用测试样本对网络进行测试。

进行工程建设,首先必须正确区分膨胀土的胀缩等级。膨胀土的胀缩等级分为强、中、弱。影响因素有五项:黏粒含量、粉粒含量、液限、塑限和缩性指数。BP 网络可以解决该问题,然而它存在内部结构黑箱化、收敛速度慢等缺点。下面使用基于 OCA 客观聚类的 RBF 神经网络解决该问题。

采集了 23 个安康胀土样本[11],其中训练样本 14 个(如表 3 所示),测试样本 9 个(如表 4 所示)。

表 3　训练样本

样本	评判指标/%					膨缩
序号	黏粒含量	粉粒含量	液限	塑限	塑性指数	等级
0	51.7	27.8	59.2	22.7	36.5	强
1	38.6	43.2	43.1	23.9	19.2	中
2	48.8	27.3	54.8	28.1	26.7	强
3	52.3	35.6	49.8	29.7	30.1	强
4	32.4	52.9	38.2	22.8	15.4	弱
5	52.5	36.8	49.6	20.3	29.3	强
6	41.8	44.3	42.8	19.7	23.1	中
7	36.9	41.6	43.4	24.1	19.3	中
8	53.2	30.5	50.7	21.9	28.8	强
9	48.9	31.7	49.3	20.7	28.6	强
10	30.8	49.7	31.7	17.1	14.6	弱
11	34.5	53.5	41.8	25.5	16.3	弱
12	49.7	33.9	52.8	26.5	26.3	强
13	39.8	46.3	47.2	23.4	23.8	中

表 4　测试样本

样本	评判指标/%					膨缩
序号	黏粒含量	粉粒含量	液限	塑限	塑性指数	等级
0	42.6	41.5	43.8	19.7	24.1	中
1	47.9	33.2	53.8	23.7	30.1	强
2	35.8	48.3	45.5	22.7	22.8	中
3	28.7	55.1	31.9	17.4	14.5	弱
4	47.2	32.3	57.4	28.1	29.3	强
5	35.3	46.9	40.8	22.2	18.6	中
6	48.2	28.6	51.6	24.2	27.4	强
7	33.6	56.3	37.8	21.6	16.2	弱
8	39.3	48.7	45.7	20.8	24.9	中

第一步:确定样本输入层节点数目为样本维数 5。输出层节点数目为 1,有 3 个取值:‘0’‘0.5’‘1’,分别代表弱、中、强。其中 0.3 以下表示‘弱’,0.3～0.7 表示‘中’,0.7 以上表示‘强’。

第二步:对所有样本进行归一化处理。令

$$x_i=\frac{Q_i-Q_{\min}}{Q_{\max}-Q_{\min}}\times a+b$$

其中,Q_i 表示样本第 i 维分量,$Q_{\min}$ 和 $Q_{\max}$ 分别表示样本第 i 维分量的最小值和最大值。取 $a=0.9$,$b=(1-a)/2=0.05$。归一化后,保留两位小数。

第三步:对归一化后的训练样本用 OCA 算法进行聚类。构造偶极子如表 5 所示。

表 5　初始偶极子

偶极子序号	0	1	2	3	4	5	6
A	1	8	4	2	6	0	3
B	7	9	11	12	13	5	10
d	0.089	0.214	0.250	0.255	0.315	0.565	1.589

由表 5 可以看出,序号为 6 的偶极子两样本之间的距离明显大于其它偶极子之间的距离,将其舍去,使其不参与下一步的聚类。

子集 A 和 B 聚类后的准则值为:

5 类时:$\varepsilon=0.667$;

4 类时:$\varepsilon=0$;

3 类时:$\varepsilon=1$;

2 类时:$\varepsilon=0$。

检测集 C 和 D 聚类后准则值为:

5 类时:$\varepsilon=0.667$;

4 类时:$\varepsilon=0.833$;

3 类时:$\varepsilon=1$;

2 类时:$\varepsilon=1$。

在聚成 4 类和 2 类时,比较子集 C 和 D 上的准则值,取较小的准则值所对应的类数。故最优方案为:分成 4 类,聚类结果,如下:

第一类:1　6　7　13;

第二类:0　5　8　9;

第三类:4　11;

第四类:2　12。

故 RBF 的隐节点数目自动确定为 4。可以看出,最优分类方案结果中没有错误的分类,即每一类都代表一个等级。计算最优聚类方案中各类的平均值,将各类的平均值作为 RBF 基函数的数据中心。然后利用 Gauss 函数计算隐层各节点输出。

第四步:使用递推最小二乘训练隐层到输出层的权值。误差限设为 0.2,即当误差小于 0.2 时,训练结束。训练 20 次,平均训练次数为 27,而使用 BP 神经网络训练,训练次数要几百甚至上千次。

使用训练好的网络对测试样本分类的结果如表 6 所示。

表6 测试样本分类结果

样本序号	期望输出	实际输出	等级
0	0.500 000	0.642 187	中
1	1.000 000	0.977 090	强
2	0.500 000	0.390 418	中
3	0.000 000	0.006 264	弱
4	1.000 000	0.876 240	强
5	0.500 000	0.305 236	中
6	1.000 000	0.967 049	强
7	0.000 000	0.091 806	弱
8	0.500 000	0.483 623	中

从上述可以看出,预测结果完全正确。基于OCA客观聚类的RBF神经网络不仅能够自适应客观地获取最佳基函数中心,并且训练速度快,正确率高。

4 结 论

本文针对影响RBF神经网络性能的关键因素是基函数中心的选取的研究,分析了当前基函数中心选取的各种方法的优缺点。由于OCA客观聚类具有自适应确定最优聚类个数的优点,通过分析RBF神经网络的结构,将OCA聚类应用于RBF神经网络中,用于自适应确定隐节点数目和基函数数据中心,克服了传统RBF神经网络不能客观确定隐节点中心的缺点。最后,本文通过膨胀土分类问题验证了该算法的性能,仿真结果表明了基于OCA客观聚类的RBF神经网络不仅具有自适应性,而且拥有训练速度快和正确高的优点。鉴于这些优点,基于OCA聚类的RBF神经网络在模式识别、数据挖掘和自动控制等领域必将有广阔的应用前景。

参考文献:

[1] 施 彦,韩立群,廉小亲. 神经网络设计方法与实例分析[M]. 北京:北京邮电大学出版社,2009:83-85.

[2] 赵清林,郭艳兵,梅 强,等. 确定RBF神经网络中心点的方法综述[J]. 广东自动化信息工程,2002(2):13-15,27.

[3] Ivakhnenko A G, Mueller J A. Parametric and nonparametric selection procedures in experimental systems analysis[J]. Systems Analysis Modelling Simulation, 1992, 9(5): 157-175.

[4] 刘 永,张立毅. BP和RBF神经网络的实现及其性能比较[J]. 电子测量技术,2007,30(4):77-80.

[5] Haykin. 神经网络原理[M]. 叶世伟等,译. 北京:机械工业出版社,2004:256.

[6] He C Z, Xu X Z. Combination of forecasts using self-organizing algorithms[J]. Journal of Forecasting, 2005, 24(4): 269-278.

[7] Ivakhnenko A G, Petukhova S A, Yudin V M. Objective selection of optimal clusterization of a data sample during compensation of non-robust random interference [J]. Journal of Automation and Information Sciences, 1993, 26(3): 45-56.

[8] 赵珩君. 基于OCA的客户细分研究[J]. 情报杂志,2009(1):8-10.

[9] 贺昌政. 自组织数据挖掘与经济预警[M]. 北京:科学出版社,2005:52-79.

[10] 贺昌政,张九龙,林 嫔. 基于数据分组处理方法的聚类分析模型[J]. 系统工程学报,2008,23(2):222-237.

[11] 吕海波,宁世朝,赵艳林,等. SOFM神经网络在膨胀土分类中的应用[J]. 土工基础,2006,20(4):90-93.

Research on RBF Neural Network Based on OCA Objective Clustering

HUANG Jing, *ZHANG Jiang*

(The School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: The key factor influencing RBF neural network performance is the selection of basis function center. Currently, there is no reliable method for selecting the center of RBF neural network. OCA objective clustering based on GMDH theory has the advantage of automatically determining the optimal clustering number. This research overcomes the disadvantage of traditional RBF network that it cannot objectively determine the number of hidden nodes by using OCA clustering in RBF neural network to determine the number of hidden nodes and the center of each radial basis function. The result of experimental simulation shows that RBF neural network based on OCA objective clustering has such advantages as adaptivity, high accuracy and fast training speed.

Key words: RBF neural network; OCA objective clustering; number of hidden nodes; basis function center

(责任编辑:陈和榜)