基于密度参数 K-均值算法的 RBF 网络及其在降水量预测中的应用

郭 皓[1], 邢贞相[1,2,3], 付 强[1,2,3], 李 晶[1]

(1. 东北农业大学 水利与建筑学院, 哈尔滨 150030; 2. 黑龙江省粮食产能提升协同创新中心, 哈尔滨 150030; 3. 黑龙江省高校节水农业重点实验室, 哈尔滨 150030)

摘 要:径向基函数(Radial Basis Funtion,简称 RBF)神经网络是一种收敛速度快、逼近能力强的前馈型神经网络。为提高网络的训练速度,采用基于密度参数的 K-均值算法,消除传统 K-均值算法对初始聚类中心的敏感性,构建了基于 K-均值算法的 RBF 降水预报模型,并应用于挠力河流域的友谊农场汛期月降水量预报中,以检验所建模型的有效性。结果表明,与标准的 K-均值算法 RBF 网络模型和 BP(Back Propagation)网络模型相比,所构建的 RBF 降水预报模型对 2008 年,2009 年,2010 年各年间汛期(6—9 月)降水量的预测平均相对误差为 9.270 7%;确定性系数为 0.96。预报精度均有所提高,且满足水文预报要求。

关键词:水文学; 降水量预测; 径向基函数神经网络; 密度参数; K-均值

中图分类号:P338 **文献标识码**:A **文章编号**:1005-3409(2014)06-0299-05

RBF Neural Network of K-Means Algorithm Based on Density Parameter and the Application to the Rainfall Forecasting

GUO Hao[1], XING Zhen-xiang[2,3], FU Qiang[1,2,3], LI Jing[1]

(1. *College of Water Conservancy & Civil Engineering, Northeast Agricultural University, Harbin* 150030, *China*; 2. *Collaborative Innovation Center of Grain Production Capacity Improvement in Heilongjiang Province, Harbin* 150030, *China*; 3. *Key Lab of Water-Saving Agriculture of Heilongjiang University, Harbin* 150030, *China*)

Abstract: The radial basis function (RBF) neural network is a feed-forward artificial neural network with high convergence speed and strong approximation capability. In order to improve the training rate of the RBF, a K-means algorithm based on density parameter was introduced to determine clustering center, which could reduce the sensitivity of traditional K-means algorithm for initial clustering centers. A rainfall forecasting model of RBF based on K-means algorithm was built, which was applied to forecasting monthly rainfall over the Youyi Farm in Naolihe catchment during the flood season, aiming to test the effectiveness of this model. The case study showed that the mean relative error of rainfall forecasting in flood season (from June to September) of the year 2008, 2009 and 2010 was 9.270 7%, and the deterministic coefficient was 0.96. It demonstrated a higher forecasting accuracy compared to a RBF model based on a standard K-means algorithm and BP (Back Propagation) model, and the rainfall forecasting results met the requirements of hydrologic prediction.

Key words: hydrology; rainfall forecasting; radial basis function neural network; density parameter; K-means

降水是自然界水循环的一个重要环节,因受多类因子的影响,降水量的变化存在复杂且多变的不确定性。高精度的降水量预测方法能提前预测降水量的变化并为径流预报提供雨情数据状态和后果,对防洪抗旱、指导生产都有重要意义[1]。降水量的预测方法众多,如回归分析[2]、灰色预测[3]、数值模拟[4]、模糊预测[5]、人工神经网络[6-7]等等。人工神经网络因具有较强的处理非线性问题的能力,无需求解具体的问

收稿日期:2014-04-15 **修回日期**:2014-05-08

资助项目:国家自然科学资助项目(51109036;51179032);教育部高等学校博士学科点专项科研基金(20112325120009);水利部公益性行业科研专项经费项目(201301096);黑龙江省级领军人才梯队后备事头人资助项目(500001);黑龙江省博士后启动金(LBH-Q12147)

作者简介:郭皓(1990—),女,黑龙江省双城人,硕士研究生,研究方向为水文预报方法研究。E-mail:ghneau@sina.cn

通信作者:付强(1973—),男,辽宁省锦州人,博导,教授,博士,研究方向为水资源分析与评价。E-mail:fuqiang0629@126.com

题解析函数，有较好的泛化能力等优点，在降水量的预测研究方法中独树一帜。目前，BP 神经网络和 RBF 神经网络是应用最多的两种网络模型。吴有训[8]等人采用遗传算法优化 BP 神经网络初始连接权值和阈值的混合算法，建立安徽宣城市汛期降水短期气候预测模型，预报误差较小。卢文喜[9]等人采用 RBF 神经网络建立桦甸市降水预测模型，结果表明，模型的后验差比值、平均绝对误差及有效系数的精度均满足要求。与 BP 网络相比，RBF 网络的隐层单元数可在训练阶段自适应调整；此外，RBF 网络输入层与隐层之间无须通过权值连接，而是采用直联的方式。由此看来，RBF 具有 BP 无法替代的优越性，可大大提高网络收敛速度，对非线性函数具有更好的一致逼近性[10]。故本文对 RBF 网络进行一定的探讨和研究，尝试用于三江平原挠力河流域降水量的预测。

1 径向基函数神经网络

径向基函数神经网络(Radial Basis Function Neural Network)是一种局部逼近的前馈式神经网络[11]，拓扑结构如图 1 所示。它模拟了人脑的局部调整、相互覆盖接受域的神经网络结构，并且能够以任意精度逼近任意连续函数。

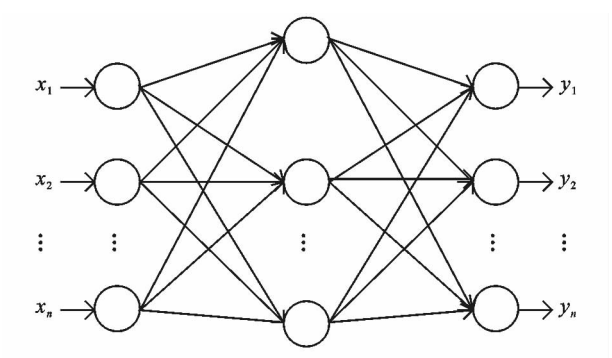

图 1 径向基网络结构图

RBF 网络的基本思想是[12]：用 RBF 作为隐单元的“基”构成隐含层空间，这样可将输入矢量直接(即不需要通过权连接)映射到隐空间。当 RBF 的中心点确定以后，映射关系便确定了。而隐含层空间到输出空间的映射是线性的，即网络的输出是隐单元输出的线性加权和。

RBF 网络中常用的径向基函数是高斯函数，因此 RBF 网络的激活函数可表示为

$$\varphi(x_p-c_i)=\exp[-\frac{1}{2\sigma^2}\|x_p-c_i\|^2] \qquad (1)$$

式中：φ——激活函数；$\|x_p-c_i\|$——欧式范数；$x_p=(x_1^p,x_2^p,\cdots,x_m^p)^T$——第 p 个输入样本；$p=1,2,\cdots,P$(P 表示样本总数)；c_i——高斯函数的中心；σ——高斯函数的方差。根据 RBF 网络的结构可得到网络输出为

$$y_j=\sum_{i=1}^{h}w_{ij}\exp[-\frac{1}{2\sigma^2}\|x_p-c_i\|^2] \qquad (2)$$

式中：w_{ij}——隐含层到输出层的连接权值；$i=1,2,\cdots,h$(h 为隐含层的结点数)；y_j——与输入样本对应的网络的第 j 个输出节点的实际输出；其他符号意义同前。

2 基于密度参数的 K-means 算法

RBF 网络学习算法需要求解的参数有三个：隐层基函数的中心、方差及隐层到输出层的权重。其中，基函数中心的选取是该网络能否成功实现的关键所在，其选取方法有多种，如随机选取中心法、自组织选取中心法、聚类分析法及正交最小二乘法。而 RBF 网络较为有效的学习算法是 K-均值(K-means)聚类法。K-means 算法基本思想[13]：它是一种基于距离的聚类算法，采用距离作为相似性的评价指标，首先在样本数据中随机选取 k 个对象作为初始的聚类中心，而后的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，即聚类中心不发生变化时为止，此时具有共同特性的样本数据便形成了特定的组群。该算法原理易于理解、计算步骤简捷、易于计算机程序的实现，但从中也不难看出 K-means 算法对初始聚类中心很敏感，聚类结果随不同的初始输入而波动，这势必影响最终各样本组群的特征。基于密度参数的 K-means 算法可以降低传统算法初始聚类中心对聚类结果波动性的影响，因此，本文采用基于密度参数 K-means 算法的 RBF 网络尝试对降水量进行预测研究。

2.1 密度参数的概念

样本数据集合：$S=\{x_1,x_2,\cdots,x_n\}$，k 个初始的聚类中心：$z_1,z_2,\cdots,z_k$。

定义 1，两个数据对象间的欧氏距离

$$d(x_i,x_j)=(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2)^{1/2} \qquad (3)$$

式中：$x_i=\{x_{i1},x_{i2},\cdots,x_{ip}\}$ 和 $x_j=\{x_{j1},x_{j2},\cdots,x_{jp}\}$——两个 p 维的数据对象。

定义 2，样本点之间的平均距离

$$\mathrm{MeanDist}=\frac{1}{C_n^2}\times\sum d(x_i,x_j) \qquad (4)$$

式中：n——样本点总数；C_n^2——n 个样本点取两个点的组合数。

定义 3，密度参数[14]空间中任一点 p 和距离 Me-

anDist,以 p 点为中心,半径为 MeanDist 的区域称为点的邻域,区域内的点的个数称为点 p 基于距离 MeanDist 的密度参数,记作 density(p,MeanDist)。

2.2　基于密度参数的 K-means 算法

在用欧氏距离作为相似性度量的 K-means 算法中,相互距离最远的 k 个数据对象比随机取的 k 个数据对象更具有代表性。但实际的数据集往往有噪声数据存在,如果只是单纯地取相互距离最远的 k 个点来代表 k 个不同的类别,有时会取到噪声点,从而影响聚类效果。一般在一个数据空间中,高密度的数据对象区域被低密度的对象区域所分割,通常认为处于低密度区域的点为噪声点。为了避免取到噪声点,取 k 个处于高密度区域的点作为初始聚类中心,具体步骤[14]:

(1) 按照式(3)计算数据对象两两之间的距离 $d(x_i,x_j)$

(2) 按照式(4)计算数据对象之间的平均距离 MeanDist。

(3) 计算全部数据对象的密度参数 density(p,MeanDist),组成一个集合 D。

(4) 从数据对象的密度参数集合中找到最大者 z_k,$z_k=\max\{\text{density}(p_i,\text{MeanDist})$

$|i\in(1,2,\cdots,n),\text{density}(p_i,\text{MeanDist})\in D\}$。如果 $d(p_i,z_k)<\text{MeanDist}$,将 density($p$,MeanDist) 从 D 中删除,z_k 为第 k 个聚类中心。

(5) 重复步骤(3),(4),直至找到 k 个聚类中心为止。

根据上述方法得到的 k 个聚类中心即为 RBF 网络最终的基函数的中心。

3　基于 RBF 网络的降水量预测模型构建

3.1　RBF 网络拓扑结构的确定

3.1.1　*输入神经元数目的选择*　输入神经元的选择对建立 RBF 网络的有效性至关重要,不合理的选择会影响网络的质量,甚至导致建模失败。因此,本文采用自相关分析技术确定 RBF 网络的输入神经元节点数。

3.1.2　*隐含层神经元数目的选择*　隐含层神经元用以存储输入层到隐含层的连接权值和阈值,能够体现训练样本与期望输出间的内在规律。目前还没有准确、科学的确定方法,多是通过经验公式和数值试验比较确定,本文也采用此方法确定,且经验公式[15]:

$$n_2=\sqrt{n_1+m}+a \tag{5}$$

$$n_2=\log 2^{n_1} \tag{6}$$

式中:n_2——隐含层神经元个数;n_1——输出层神经元个数;m——输入神经元个数,a——[0,1]之间的常数。

3.2　降水量预测模型的构建

网络模型的创建采用 Matlab 的 RBF 工具箱来实现,具体的程序语言为:net = newrb(P,T,goal,spread,mn,df),其中,P 和 T 分别为样本的输入向量和输出向量;goal 为均方误差,本文取 goal=0.001,目的是防止过度拟合;spread 为径向基函数的宽度,本文通过隐层神经元各个中心之间的距离来确定,其计算公式

$$\text{spread}=b\times d_i \tag{7}$$

式中:b——重叠系数,一般取大于 1 的整数[16];d_i——隐含层基函数中心间的最小距离,基函数中心由基于密度参数的 K-means 算法求得。

4　实例研究

4.1　试验数据

现以三江平原友谊农场 1956—2007 年各年汛期(6—9 月)降水量资料为例,建立基于密度参数 K-means 算法的 RBF 神经网络预测模型。选取 1956—2005 年各年汛期降水数据用于网络训练,2006 年和 2007 年汛期降水数据用于网络检验,并对 2008,2009,2010 年汛期各月降水量进行预测。

4.2　数据预处理

由于 RBF 网络对输入、输出数据的要求在[0,1]之间,故在训练之前先将降水数据按以下公式进行归一化处理,即

$$Y_i=\frac{X_i-X_{\min}}{X_{\max}-X_{\min}} \tag{8}$$

式中:X_i,Y_i——归一化处理前、后的降水量;$X_{\min}$,$X_{\max}$——所有降水量的最小值和最大值。经过以上处理后的数据将作为 RBF 神经网络模型的样本。

4.3　RBF 网络结构的确定及训练

根据降水量的特性,设计 RBF 神经网络,主要是确定输入层与隐含层的节点数。对于友谊农场汛期降水量序列进行相关分析(图 2,图 3 分别给出降水量序列的自相关函数图和偏相关函数图),据自相关图的拖尾性和偏相关图的截尾性,可判别降水量序列适用于 AR 模型,并通过 AIC 准则[17]分析得到马尔科夫过程的阶数应为 4(表 1 中给出 AIC 准则的计算结果),即 RBF 神经网络中输入层的节点个数为 4。隐含层神经元用以存储输入层到隐含层的连接权值,体现训练样本与期望样本间的内在规律。根据经验公式和试算,隐含层节点数选定 6 个,至此,本文网络拓扑结构为 4∶6∶1。

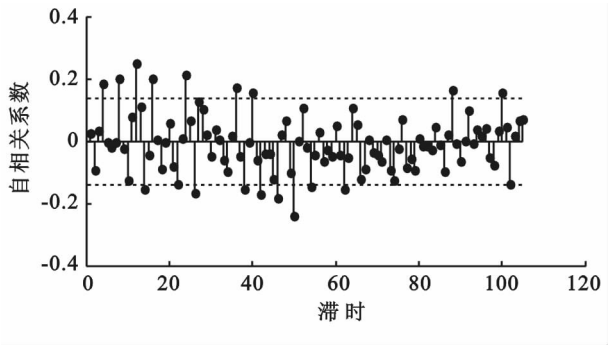

图2 降水量序列自相关函数

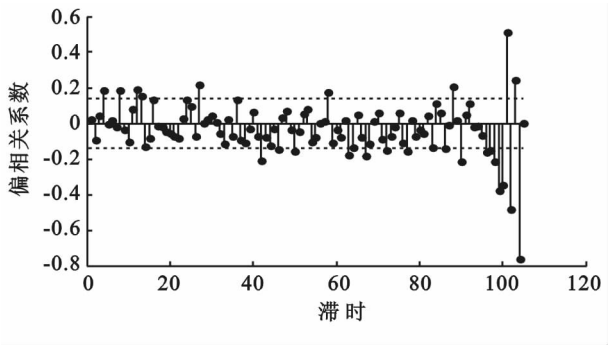

图3 降水量序列偏相关函数

表1 AIC准则计算结果

阶数	AIC值
4	1622.4
8	1688.9
12	1632.5

利用已构建的神经网络对1956—2005年的汛期降水量数据训练188次达到收敛,网络拟合误差达到0.001,其网络误差变化图及网络训练拟合图(如图4,图5),通过拟合曲线看出,训练阶段计算降水量变化趋势与实测降水量趋势拟合度较高。为验证所构建RBF网络模型的泛化能力,对2006年和2007年汛期各月降水量进行检验计算,验证期计算结果如表2所示,平均预报相对误差为9.666 2%,确定性系数为0.98。经以上验证,可利用已构建的RBF网络进行挠力河流域降水量预测,表3给出了模型对2008,2009及2010年汛期各月降水量的计算结果(为便于对比,表中还列出了标准K-Means算法RBF网络与BP网络的计算结果)。

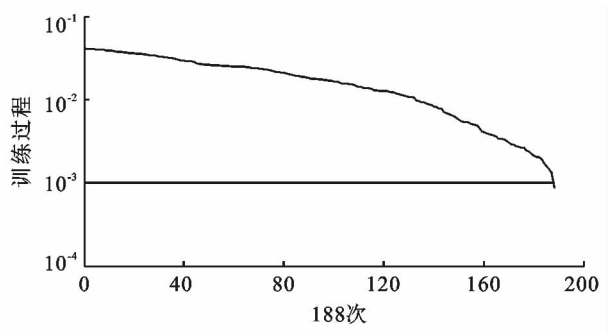

图4 RBF神经网络拟合误差变化

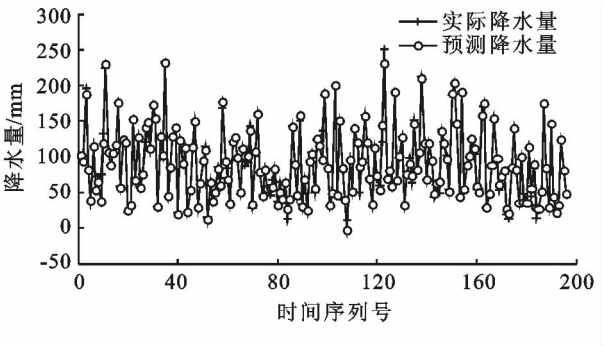

图5 RBF神经网络训练拟合

表2 验证期(2006—2007年)汛期各月降水量计算结果及分析

时间	实际降水量/mm	预测降水量/mm	相对误差/%
2006-06	107.5	99.4932	−7.4482
2006-07	100.8	95.6321	−5.1269
2006-08	82.3	77.8052	−5.4615
2006-09	27.4	30.8841	12.7157
2007-06	17.7	15.6916	−11.3469
2007-07	33.6	37.5182	11.6613
2007-08	139.5	125.7038	−9.8897
2007-09	29.7	25.6372	−13.6795
平均误差	—	—	9.6662
确定性系数	—	0.98	—

4.4 计算结果与分析

表3中计算结果表明,应用已建立的RBF网络模型对2008,2009及2010年汛期各月降水量预测的平均相对误差为9.270 7%,确定性系数为0.96;而标准K-means算法预测的平均相对误差为12.894 0%,确定性系数为0.92;BP网络模型的平均相对误差为11.242 0%,确定性系数为0.94。与此同时,已建立的RBF网络模型训练拟合时间及收敛次数也有所提高,分别为85 s及188次;而标准K-means算法的训练拟合时间及收敛次数分别为100 s及266次;BP网络模型的训练拟合时间及收敛次数分别为121 s及420次。由此可见,与标准K-means算法的RBF模型和BP模型相比,本文所构建的RBF网络计算速度分别提高15%,30%;预报平均相对误差分别降低28%,19%;确定性系数均相当。

5 结论

K-means算法是RBF神经网络常用的学习算法,但是传统的K-means由于随机选取初始聚类中心不同而造成聚类结果的波动,从而导致网络产生不同的收敛能力。本文应用基于密度参数寻找初始聚类中心的思想,尝试消除随机选取聚类中心的不稳定

性和敏感性，并利用基于密度参数的 K-means 算法求得 RBF 网络基函数的宽度，构建了友谊农场基于密度参数 K-Means 的 RBF 降水量预测模型。通过实例研究，得出以下结论：

表 3　改进 RBF 网络与标准 K-means RBF 网络及 BP 网络降水量预测结果对比

时间	实际降水量/mm	改进的 RBF 网络		标准 K-means 的 RBF 网络		BP 网络	
		预测降水量/mm	相对误差/%	预测降水量/mm	相对误差/%	预测降水量/mm	相对误差/%
2008-06	29.0	25.0815	−13.5121	23.9715	−17.3397	24.8602	−14.2752
2008-07	112.4	105.0493	−6.5398	100.1356	−10.9114	103.2024	−8.1829
2008-08	37.5	40.0862	6.8965	42.0483	12.1288	40.8573	8.9528
2008-09	41.7	36.0726	−13.4950	34.3291	−17.6760	33.8621	−18.7959
2009-06	96.3	86.854	−9.8086	85.6432	−11.0663	90.0321	−6.5087
2009-07	142.8	145.2525	1.7174	152.4741	6.7746	159.9773	12.0289
2009-08	124.9	108.9841	−12.7429	105.0331	−15.9062	104.5027	−16.3309
2009-09	97.9	100.0613	2.2077	101.1410	3.3105	100.5302	2.6866
2010-06	53.6	43.6932	−18.4828	45.0775	−15.9002	44.8511	−16.3226
2010-07	174.5	160.5336	−8.0037	152.6580	−12.5169	159.0021	−8.8813
2010-08	148.5	133.4076	−10.1632	125.8915	−15.2246	130.5544	−12.0846
2010-09	39.2	42.2099	7.6783	45.4612	15.9724	35.3375	−9.8533
平均误差	—	—	9.2707	—	12.8940	—	11.2420
确定性系数	—	0.96	—	0.92	—	0.94	—

注：表中改进的 RBF 网络为本文构建的 RBF 网络。

（1）从网络收敛速度来看，基于密度参数 K-means 算法的 RBF 网络模型与传统 K-means 算法的 RBF 网络相比，对聚类数、基函数中心的确定等方面有着一定的优越性，收敛速度、预报精度均有明显的提升。

（2）对于挠力河流域友谊农场汛期月降水量的预测而言，改进的 RBF 网络模型与 BP 网络模型相比，在精度上有较大的提高，2008，2009，2010 年汛期月降水量的平均相对误差分别降低 19%，29% 及 6%，确定性系数分别提高 3%，4% 及 2%，为在不确定因素较强的降水量预测提供了一种新方法；

（3）应用本文所述的 RBF 网络模型对月降水量进行预测时，网络的节点数采用经验公式和网络试算，在日后的研究中，可采用智能优化算法优化 RBF 网络的隐层节点数，以进一步提高模型的计算精度。

参考文献：

[1] 常青，赵晓莉. 时间序列模型在降水量预测中的应用研究[J]. 计算机仿真，2011，28(7)：204-206.

[2] 唐亚松，张鑫，蔡焕杰，等. 一种基于回归分析与时序分析的降水预报模型[J]. 水土保持通报，2009，29(1). 88-92.

[3] 李才媛，顾永刚. 灰色预测模型在长江上游流域面雨量预报中的应用[J]. 气象科技，2004，31(4)：223-225.

[4] 钱莉，杨晓玲，殷玉春，等. 最优子集回归在武威市降水预报中的应用[J]. 干旱区研究，2009，26(6)：895-900.

[5] 孙才志，林学钰. 降水预测的模糊权马尔可夫模型及应用[J]. 系统工程学报，2003，18(4)：294-299.

[6] 封毅，武博强，崔灵周. 基于 BP 神经网络的台风降雨量预测研究[J]. 水土保持研究，2012，19(3)：289-293.

[7] 葛彩莲，蔡焕杰，王健，等. 基于 BP 神经网络的降雨量预测研究[J]. 节水灌溉，2010(11)：7-10.

[8] 吴有训，王周青，汪文烈，等. 遗传算法优化 BP 网络的汛期降水预测模型[J]. 安徽农业大学学报，2013，40(2)：299-303.

[9] 卢文喜，杨磊磊，杨忠平，等. 逐步回归时间序列和 RBF-ANN 在降水预测中的应用[J]. 重庆大学学报：自然科学版，2012，35(11)：131-135.

[10] 陈玉红. RBF 网络在时间序列预测中的应用研究[D]. 哈尔滨：哈尔滨工程大学，2009.

[11] 伊燕平，卢文喜，张耘，等. 基于径向基函数神经网络的地下水数值模拟模型的替代模型研究[J]. 水土保持研究，2012，19(4)：265-269.

[12] 张德丰. MATLAB 神经网络应用设计[M]. 北京：机械工业出版社，2009.

[13] 靳玉萍，党婕. 基于径向基神经网络改进算法优化锅炉燃烧效率[J]. 计算机应用，2013，33(06)：1771-1779.

[14] 张建辉. K-means 聚类算法研究及应用[D]. 武汉：武汉理工大学，2007.

[15] 胡可云，田凤占，黄厚宽. 数据挖掘理论与应用[M]. 北京：清华大学出版社，2008.

[16] 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.

[17] 王文圣，丁晶，金菊良. 随机水文学[M]. 北京：中国水利水电出版社，2008.