DOI:10.13878/j.cnki.jnuist.2020.04.009

陈晓平[1]　陈易旺[1]　施建华[2]

基于机器学习的降雨量雷达回波数据建模与预测

摘要

以浙江省2016年1—10月的雷达回波强度数据为基础,分别应用随机森林模型、BP神经网络模型、卷积神经网络模型来预测降雨量并进行对比.建模分析结果表明,随机森林模型预测效果精确度较低,容易低估较大的降雨强度,而BP神经网络和卷积神经网络预测的效果都比随机森林好,特别是卷积神经网络,其预测值与真实值更加接近,且对较大的降雨强度拟合较好.

关键词

降雨量;BP神经网络;卷积神经网络;随机森林

中图分类号 O212;P412.13

文献标志码 A

收稿日期 2019-03-07

资助项目 国家自然科学基金(11601083);福建师范大学创新团队基金(IRTL1704);福建省高等学校科技创新团队培育计划(IRT-STFJ);福建师范大学研究生教育教学改革研究项目资助;数字福建气象大数据研究所项目;福建省数据科学与统计重点实验室开放课题(2020L0704,2020L0703)

作者简介

陈晓平,男,博士,副教授,研究方向为统计学理论与方法.xpchen@fjnu.edu.cn

施建华(通信作者),男,博士,副教授,研究方向为统计理论与方法.v0085@126.com

1 福建师范大学 数学与信息学院,福州,350117

2 闽南师范大学 数学与统计学院,漳州,363000

0 引言

降雨对人类生活、国民经济起着重要的作用,降雨量变化造成洪涝、干旱等极端情况,对农业、水资源、生态环境等存在很大影响.准确的降雨信息对于水资源的规划和管理至关重要,也是水库抗旱和防洪的关键.然而,由于产生降雨的大气过程的复杂性以及在空间和时间上各种尺度的巨大变化,造成了降雨的预测具有很大的挑战性.随着气象卫星以及天气雷达等先进设备、技术的发展,人们在天气预报方面取得了许多进步,但是要获得准确的降雨预报仍然面临着很大的问题.降雨量具有非线性、复杂性、多样性和不稳定性等特点,且受诸多因素的影响,而数据采集方面,随着近年来科技的发展,卫星和天气雷达每年提供PB级气象数据与以往的数据有着显著的、本质上的差异,故用传统的技术、方法预测降雨存在着模糊性和不确定性,预测难度大增,往往无法取得很好的预测效果.

另一方面,随着信息技术和计算科学的迅猛发展,计算机计算能力得到大幅度提高,统计机器学习方法被广泛应用于各个领域[1-2].该方法具有高度的非线性、灵活性和数据驱动学习能力,可以应用在降雨量的预测中,也可以得到比传统方法更好的降雨量预测结果.因此,通过统计机器学习方法分析气象数据,发现其潜在的规律,从而更准确地预测未来的降雨量,是一个很有意义的课题.

Baik和Hwang基于北太平洋西部热带气旋的31年样本,利用多元线性回归方法和BP神经网络对未来12、24、36、48、60和72 h的气旋强度进行预测[3].结果表明,除了对未来12 h的预测误差相近,BP神经网络模型解释的方差百分比在其他所有时间间隔内均大于回归模型解释的方差百分比,BP神经网络模型对未来其他时刻的预测误差比回归模型小10%~16%,显示了BP神经网络在热带气旋强度预报中的应用潜力.Guhathakurta[4]首次在利用神经网络技术进行季风降水预报时引入了尺度化的思想,他分析了36个气象分区的月降雨量时间序列数据,所用的模型较好地捕捉了输入-输出的非线性关系,较准确地预测了独立周期内的季节降水.所有印度季风降水预报都是利用各分区的面积加权降水预报生成的,结果表明,向上尺度有助于更好地捕捉全印度降水的变化,有助于预测极端年份的降雨量,比基于单一时间序列建立的印度所有降雨量的神经网络模型都要好.徐晓岭等[5]给出了全样本场合下卡帕分布参数的矩估计,估计北京、天津、

南京、上海、广州的月降水量.崔玫意等[6]基于1951—2010年河北省21个气象站逐日降水观测资料,拟合逐年日最大降水量序列,借助K-S与A-D方法进行拟合优度的比较.Sulaiman等[7]利用人工神经网络对月强降水进行预报,为此收集并使用了1965—2015年地方气象站的降水数据,利用以往降水值的不同组合作为预测输入,利用均方误差和相关系数对人工神经网络模型的性能与ARIMA模型进行了比较.结果表明,该人工神经网络模型能够较好地预测强降水事件.

在技术设备方面,雷达通过发射电磁波后收到的反馈,获得物体到雷达的距离、方位角和物体当前的径向速度、高度等[8],相对于地面测量的优势在于其覆盖范围广、穿透强,基本不会受温度、风等外部因素的影响,因此在各领域被广泛应用.例如,在军事领域上可以探测复杂的地形,在环保领域上可以监测空气质量,在地质领域上可以勘探石油煤炭等.天气雷达每年提供PB级气象数据,数据量大,传统的统计学方法预测降雨往往无法取得很好的效果.

本文以浙江省气象站的降雨数据(逐小时)以及雷达回波数据(逐10 min)为基础,对其筛选、整理、预处理,分别采用随机森林、BP神经网络(BPNN)和卷积神经网络(CNN)等建模方法对未来1~2 h的降雨量进行预测.

1 数据样本及其处理

1.1 数据说明

本文使用的雷达回波数据为浙江省雷达站2016年1—10月反射率因子资料,降雨量数据为浙江气象站的降雨资料.雷达回波数据如图1所示(纵、横坐标都是像素,6幅分图是同一区域每隔10 min的雷达回波图),其中心点为气象站所在的位置,高度为1.5 km,时间分辨率为10 min,大小为100像素×100像素,回波强度范围为0到70 dBz.降雨量数据的时间分辨率为1 h.对于每个时刻的降雨量预测,输入变量为前一个小时内的6张雷达图数据.通过池化操作[9]和风向法(见下面介绍)把原始的100×100的回波强度雷达数据变为17×17的数据,为后面建立的预测模型提供了数据支持.

1.2 风向法数据处理

首先对每一张100×100的回波强度雷达图,利用池化操作将其转化为25×25的数据,然后利用前一时刻的数据和当前时刻的数据计算出风向,最后利用风向从25×25的数据中选取出17×17的数据.

1)平均池化

对每张100×100的回波强度雷达图,以步长为2,对每个2×2区域都进行平均,最后得到50×50的数据,如图2所示.

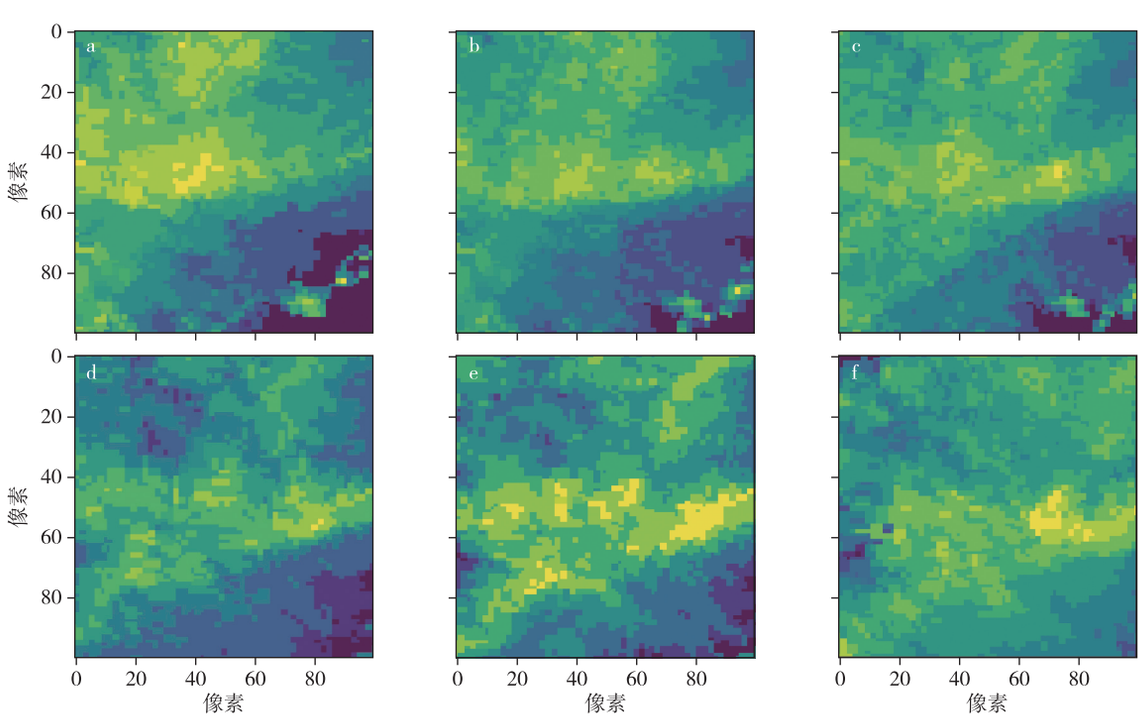

图1 1 h内逐10 min的雷达回波数据

Fig. 1 Radar echo data for 10 minutes in one hour

2)最大池化

对平均池化后得到的 50×50 的数据,以步长为 2,对每个 2×2 区域都取最大值,最后得到 25×25 的数据,如图 3 所示.

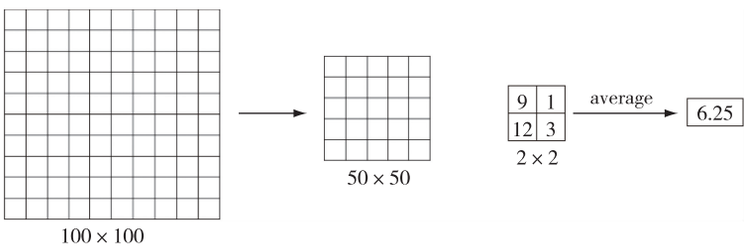

图 2 平均池化示意图

Fig. 2 Average pooling schematic

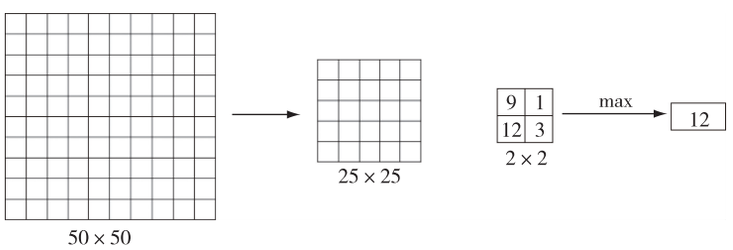

图 3 最大池化示意图

Fig. 3 Maximum pooling schematic

3)计算风向

首先对上面得到的 25×25 的数据,计算其中最大 5 个数的平均位置.然后根据前一时刻的位置和当前时刻的位置决定风向,可能的风向有 12 个,分别是:西—东、西—南、西—北、东—西、东—南、东—北、北—南、北—西、北—东、南—北、南—西、南—东.

4)选取区域

直观上,阴雨云会向着风的方向运动,这就意味着根据风向选取区域会使预测更加精准,因此本文从 25×25 的数据抽取出 17×17 的数据.

对于不同的风向,抽取不同的区域.若风向是西—东,则选择[4:21,0:17]的区域;若风向是西—南,则选择[6:23,0:17]的区域;若风向是西—北,则选择[2:19,0:17]的区域;若风向是东—西,则选择[4:21,8:25]的区域;若风向是东—南,则选择[6:23,8:25]的区域;若风向是东—北,则选择[2:19,8:25]的区域;若风向是北—南,则选择[0:17,4:21]的区域;若风向是北—西,则选择[0:17,2:19]的区域;若风向是北—东,则选择[0:17,6:23]的区域;若风向是南—北,则选择[8:25,4:21]的区域;若风向是南—西,则选择[8:25,2:19]的区域;若风向是南—东,则选择[8:25,6:23]的区域.

2 研究的理论与方法

2.1 随机森林模型

随机森林模型是由多个相互独立的 CART 决策树结合而成的建模方法,该模型既能被用来解决分类问题,也能被用来解决回归问题[10],能在运算量没有显著提高的前提下提高预测精度.若待预测的变量为类别变量,则随机森林的最终结果由所有 CART 决策树投票决定;若待预测的变量为数值变量,则随机森林的最终结果是所有 CART 决策树的平均值[11].其中每棵 CART 决策树的训练数据是由自助法(bootstrap)获得的,也就是从原始数据集中有放回地重复随机抽取数据放入训练数据集中,因此,每棵决策树的训练数据各不相同[12].随机森林算法原理框架如图 4 所示.

2.2 BP 神经网络模型

BP 神经网络[13]是一种按误差反向传播(简称误差反传)训练的多层前馈网络,其基本思想是梯度下降法,利用梯度搜索技术,以使得网络的实际输出值和期望输出值的误差均方差为最小.而其逆向传播算法的基本原理是通过迭代处理训练元组的数据

集,将每一个元组经过神经网络模型处理后的输出结果与训练集中已经给定的响应变量值进行比较,并计算误差,从而根据误差对每一层的权重与偏置项进行调整.上述过程循环进行,直到满足停止条件为止.

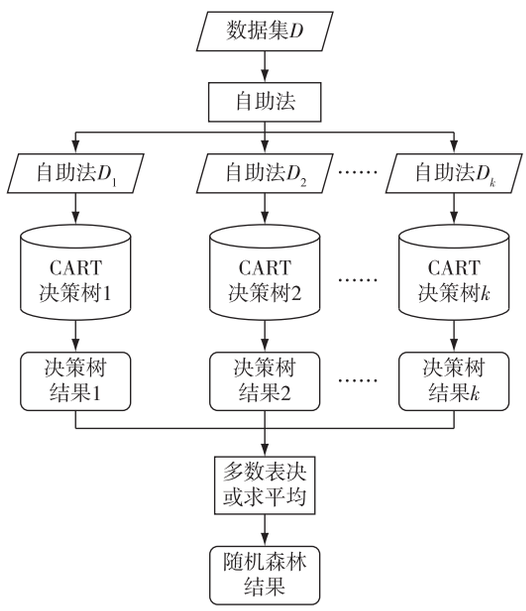

图4　随机森林原理

Fig. 4　Principle of random forest model

假设训练数据集为 $\{(x(1),y(1)),(x(2),y(2)),\cdots,(x(m),y(m))\}$,采用批量更新的方法,这 m 个数据的相应总误差为

$$L=\frac{1}{m}\sum_{i=1}^{m}L(i),$$

其中 $L(i)$ 为单个样本的误差,其定义如下:

$$L(i)=\frac{1}{2}\sum_{k=1}^{n}(d_k(i)-y_k(i))^2,$$

其中 $d_k(i)$ 为样本 i 输出层节点 k 的输出,而 $y_k(i)$ 为其真实值.因此有

$$L=\frac{1}{2m}\sum_{i=1}^{m}\sum_{k=1}^{n}(d_k(i)-y_k(i))^2.$$

误差逆向传播的每次迭代都是沿着误差相对权重值的负梯度方向来更新:

$$W_{nj}^{(l)}\leqslant W_{nj}^{(l)}-\eta\frac{\partial L}{\partial W_{nj}^{(l)}},$$

其中,$W_{nj}^{(l)}$ 表示第 l 层的第 n 个节点关于第 $l-1$ 层第 j 个节点的权重,η 为学习率.

由于BP神经网络的学习能力太强,容易产生过拟合问题.可以采取两种方法来解决过拟合问题:

1)早停.通过对训练数据集进行训练,学习调整各个权重和偏置项,将验证数据集输入模型计算误差,如果验证数据集的误差随着训练数据集误差的降低反而升高,那么就停止训练,返回此时权重和偏置项.

2)正则化.其基本思想是在误差函数中加入反映模型复杂程度的指标,使得模型不要任意拟合训练数据中的噪声.

常用的指标有两种,一种是 L_1 范数,其中 w_i 表示权重:

$$R(W)=\|w\|_1=\sum_i|w_i|.$$

另一种是 L_2 范数:

$$R(W)=\|w\|_2^2=\sum_i|w_i^2|.$$

2.3　卷积神经网络模型

卷积神经网络(CNN)是一种前馈神经网络,人工神经元可以响应周围单元,可以进行大型图像处理,其主要是由输入层、卷积层、池化层、全连接层和输出层构成的,卷积层和池化层是特征提取的关键,并且在前几层中交替出现[14].

卷积神经网络具有如下3个优势特性.

1)局部感知

1962年,Hubel等发现猫脑皮层的神经元有局部感知的特点,从而提出了感知野的概念[15].1982年,Fukushima等基于感知野建立了神经认知机模型[16],Lecun等[17]受其启发建立了有局部连接特性的卷积神经网络,其中所提到的局部连接指的是下一层中的每个节点都只与当前层的部分节点相连,从而大幅度地减少了权重的个数.

2)权重共享

虽然通过局部感知能大幅度地减少权重的个数,但权重个数仍然过多,于是权重共享的方法[18]被提出,其原理是:从大尺寸图像中选取一小部分,从这部分中学习到一些有用的特征,这样就可以把这个操作在大尺寸图像中的任何地方使用.

卷积神经网络利用卷积操作实现局部感知和权值共享,而卷积操作则利用卷积核实现,如图5所示,其计算公式如下:

$$y_i=f\Big(\sum_{i=1}^{n}W_i x_i+b\Big),$$

其中,f 是激活函数,W_i 是卷积核的权重值,b 是偏置项.

但单个卷积核只能学习到一种特征,因此,卷积神经网络中每一个卷积层都会有多个卷积核,从而

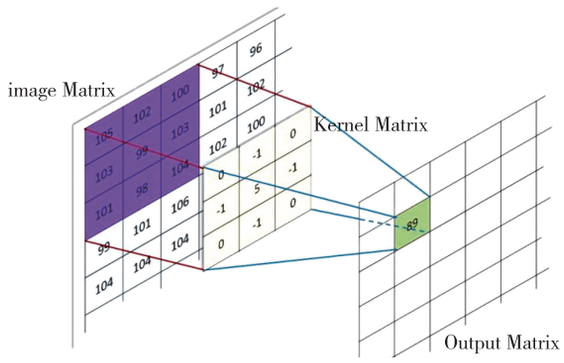

图 5　卷积操作过程

Fig. 5　Convolution operation

充分提取多种特征.

3)下采样

实现局部连接和权重共享后,即使卷积神经网络的权重数量已经降低到合理范围内,但卷积神经网络的特征矩阵往往还会出现过大问题,这不仅导致计算量增大,还容易造成过拟合.为此,在卷积神经网络的基础上又提出了下采样方法.它一般在卷积层之后,对卷积层的输出分别通过平均池化或最大池化法进行统计,也就是计算平均值或者最大值这两种方式进行统计,所以下采样又称为池化.图 6 显示了 2×2 的最大池化的过程.

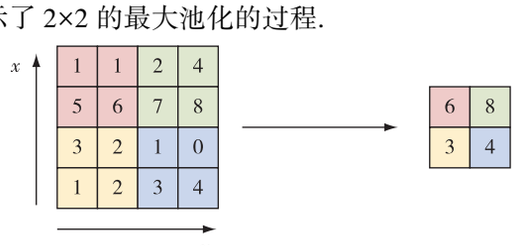

图 6　最大池化过程

Fig. 6　Maximum pooling process

3　预测模型

3.1　随机森林建模分析

我们首先选取原始雷达回波数据和降雨量数据的前 90%作为训练数据集,后 10%作为测试数据集.通过控制变量的方法,在分裂节点抽取的特征数为总特征数的 10%~90%、CART 决策树的棵数从 0 到 100 变化时,观察模型均方误差的变化情况.结果如图 7—11 所示,图中横坐标为 CART 决策树的棵数,纵坐标为模型的均方误差.

从图 7—11 中可以看出,对于不同的特征数,模型的均方误差都是随着 CART 决策树的增加先降低而后趋于平稳.其中,从图 7 可以看到,当分裂节点抽取的特征数为总特征数的 10%时,CART 决策树的棵数取 25 可以使模型的均方误差达到最小,最小值为 1.78;从图 8 可以看到,当分裂节点抽取的特征数为总特征数的 30%时,CART 决策树的棵数取 15 可以使模型的均方误差达到最小,最小值为 1.82;从图 9 可以看到,当分裂节点抽取的特征数为总特征数的 50%时,CART 决策树的棵数取 15 可以使模型的均方误差达到最小,最小值为 1.89;从图 10 可以看到,当分裂节点抽取的特征数为总特征数的 70%时,CART 决策树的棵数取 97 可以使模型的均方误差达到最小,最小值为 1. 89;从图 11 可以看到,当分裂节点抽取的特征数为总特征数的 90%时,CART 决策树的棵数取 55 可以使模型的均方误差达到最小,最小值为 1.93.经过综合,本文随机森林模型中节点分裂时随机抽取特征数为总特征数的 10%,而模型规模即 CART 决策树的棵数为 25.

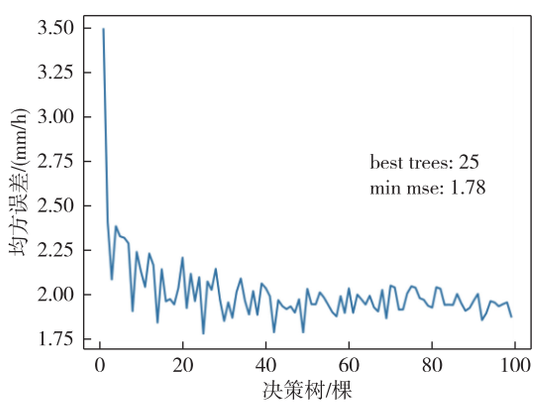

图 7　10%特征数

Fig. 7　MSE under 10% characteristic number

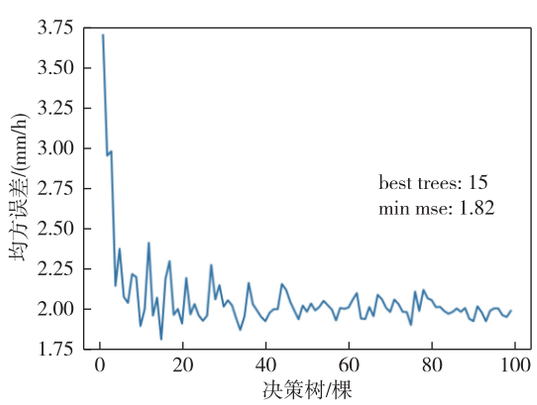

图 8　30%特征数

Fig. 8　MSE under 30% characteristic number

使用训练数据构建完模型后,还需要评估模型

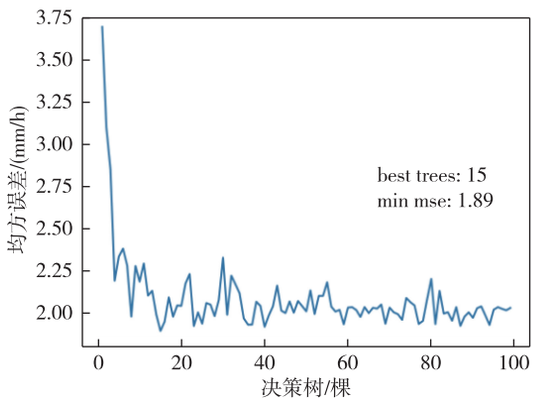

图 9　50%特征数

Fig. 9　MSE under 50% characteristic number

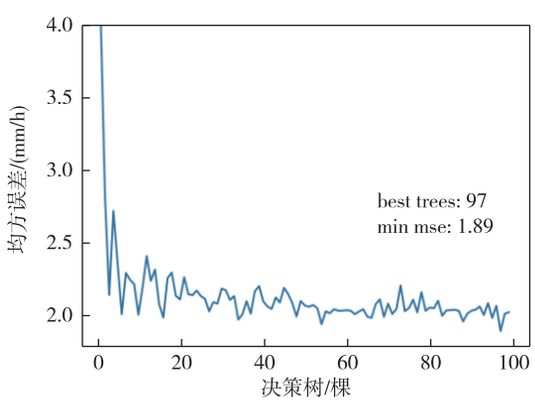

图 10　70%特征数

Fig. 10　MSE under 70% characteristic number

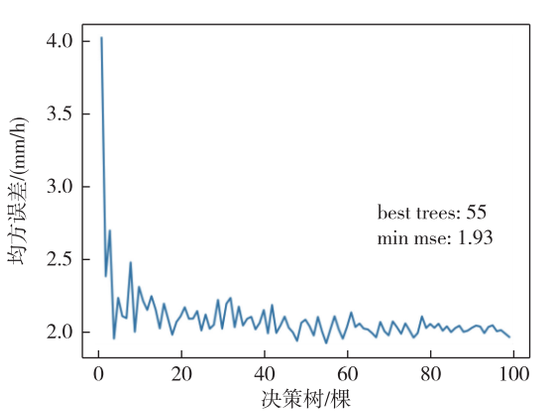

图 11　90%特征数

Fig. 11　MSE under 90% characteristic number

的泛化能力.验证模型泛化能力主要是将模型作用于测试数据集,比较模型预测值与测试数据实际值之间的差异,差异越小则说明模型泛化能力越好.我们将构建好的随机森林模型应用于测试数据集,得到模型预测结果与测试数据实际值的散点图以及残差图分别如图 12 和图 13 所示.从图中可以看出,随机森林方法虽然对低降雨强度拟合较好,但容易低估较大的降雨强度.

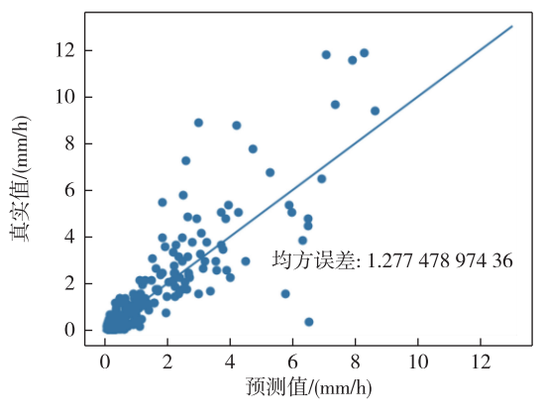

图 12　降雨量真实值与预测值的散点图

Fig. 12　Scatter of actual rainfall and random forest predicted values

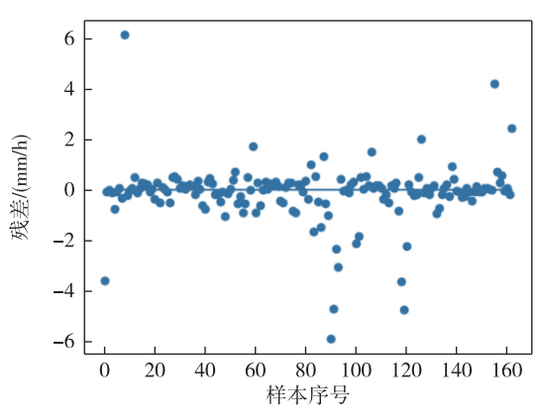

图 13　残差图

Fig. 13　Residual distribution of random forest prediction

3.2　BP 神经网络建模分析

3.2.1　数据归一化

数据在输入 BP 神经网络之前必须要进行数据归一化,也就是将数据映射到[0,1]区间或更小的区间.本文采用最小最大法将数据映射到[0,1]区间,转换函数的定义如下:

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}.$$

3.2.2　BP 神经网络设计

1)网络层数的设计

具有单隐含层的 BP 神经网络可以拟合任意函数,增加隐藏层虽可进一步降低误差,但随着层数的增加,会导致梯度消失和梯度爆炸的问题,模型反而更易得到局部最优解.因此经多次试验后选用包含两个隐含层结构的 BP 神经网络.

2)输入层和输出层的设计

输入层以及输出层的节点数分别由样本、响应变量的特征个数决定.本文的目标是对气象站未来 1 h 累计降雨量进行预测,所涉及的样本、响应变量的特征个数分别为 17×17×6=1 774 个以及 1 个,所以输入层、输出层的节点数分别为 1 774 及 1.

3)隐含层节点数目设计

隐含层最佳节点数的确定是一个难点,若隐含层节点数过少,则会导致欠拟合问题的出现;若隐含层节点数过多,则很有可能出现过拟合问题,且训练时间大大增加.所以本文考虑在一定的范围内,先训练包含较少隐含层节点的 BP 神经网络,然后逐渐增加,当训练误差达到最小时对应的节点数就是最佳的节点数.本文对 1~50 个隐藏层节点的神经网络进行均方误差比较,具体结果如图 14 所示,训练数据和测试数据的均方误差都是先下降然后趋于平稳,说明模型性能良好.可以看出,当节点数大于 32 时,测试数据的均方误差开始趋于平稳,因此我们将隐藏层节点设置为 32.

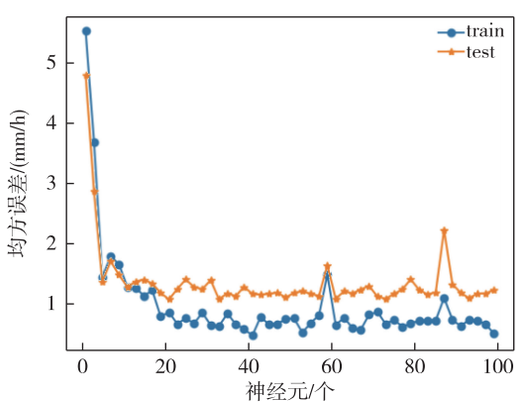

图 14 不同隐含层节点数下的模型均方误差

Fig. 14 MSE of the BPNN modeling under different hidden layer nodes

3.2.3 初始化参数的选取

1)初始权重值和偏置项的选取

由于 BP 神经网络具有高度非线性,导致其误差曲面是非凸的,包含局部极小值点,故初始权重值和偏置项的选取要在零点左右.且初始权重值和偏置项的选取还决定了模型的初始训练误差及其之后的变化.因此,本文限制初始权重值和偏置项在标准正态分布的 2 倍标准差之内.

2)学习次数

合适的学习次数能使预测的精度更高,但是一味地增加学习次数只会使预测的精度降低.本文先设置模型的最优隐含层节点数为 32,然后设置学习次数从 1~500,结果如图 15 所示.

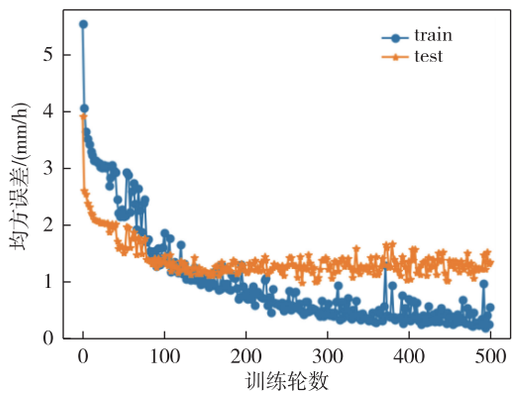

图 15 不同学习次数下的模型均方误差

Fig. 15 MSE of the BPNN modeling under different learning times

从图 15 中可以看出,神经网络模型训练集数据下的均方误差随模型学习次数增多而减小,而模型在测试数据集下的均方误差并没有随着学习次数增加而一直下降,而是先下降然后缓慢增加,这是由于过拟合导致的.由于均方误差在 170 次后开始缓慢增加,因此将 BP 神经网络的学习次数设为 170.

3)期望误差的选取

所谓期望误差,是根据当前实际情况对训练误差给定的阈值,若训练误差达到了这个值就停止训练.这是由于,若是给定的误差过小使得神经网络难以达到,就会导致训练次数过多,发生过拟合现象;若是给定的误差过大则会使得神经网络过早地停止学习,难以达到最高的精度.通过多次的实验观察,我们将期望误差设定为 1.1.

4)学习步长(速率)与梯度下降方法

由于学习步长与梯度下降方法不仅影响着各个权重和偏置项的变化,同时影响着 BP 神经网络的收敛速度,所以学习步长与梯度下降方法的选择也尤为重要.学习步长设定过大,则可能会导致 BP 神经网络不稳定;而学习步长设定过小,虽能避免出现网络不稳定的问题,但会导致训练时间过长,甚至出现不能收敛的问题.对于复杂的 BP 神经网络,好的梯度下降方法应该可以自适应地设置学习步长,加速收敛进度,避免落入局部极小值.根据多次实验,本文选择 Adam 梯度下降方法,并将学习步长设定为 0.000 5.

经过上述多次实验,对 BP 神经网络的参数进行

调整,最终确定模型的隐含层节点个数、学习次数分别为 32、170,同时将初始权重值和偏置项限制在标准正态分布的 2 倍标准差之内,并将期望误差、学习速率分别确定为 1.1、0.000 5,而选 Adam 方法为梯度下降方法,选择 Relu 函数作为激活函数.据此构建 BP 神经网络预测模型,得到图 16—18.

从图 16 可以看到,测试数据的均方误差随着训练数据均方误差的降低而降低,表明该模型相当理想.图 17 显示了降雨量真实值与预测值的散点图,降雨量真实值与预测值形成的散点基本分布在 $y=x$ 周围,最终降雨量真实值与预测值的均方误差为 1.16,表明模型拟合效果较好.结合散点图和残差图,与随机森林模型相比,BP 神经网络模型对较大的降雨强度拟合得更好.

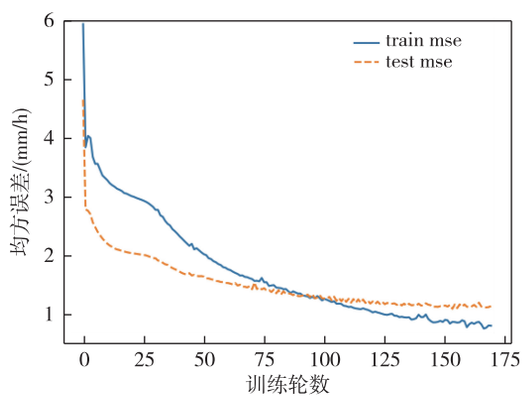

图 16　迭代过程中训练集和测试集均方误差的变化

Fig. 16　MSE variation of BPNN training set and test set during iteration

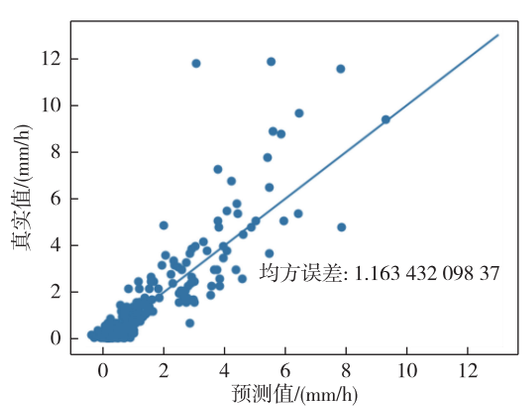

图 17　降雨量真实值与预测值的散点图

Fig. 17　Scatter of actual rainfall and BPNN predicted values

3.3　卷积神经网络建模分析

依据卷积神经网络原理,可以设计如图 19 所示的降雨量预测卷积神经网络结构,它包括了一个输入层、两个卷积层、两个池化层、两个全连接层和一个输出层.

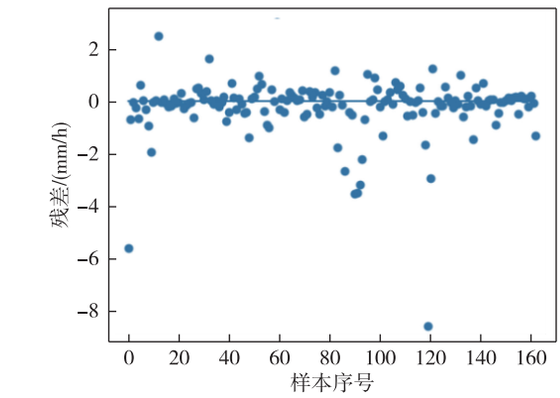

图 18　残差图

Fig. 18　Residual distribution of BPNN prediction

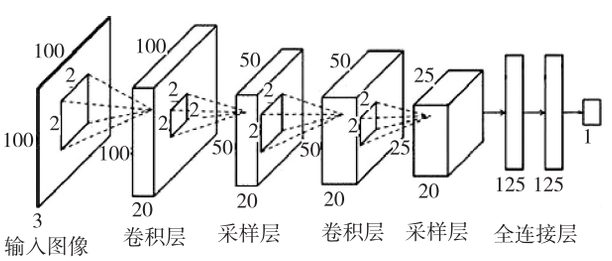

图 19　降雨量预测卷积神经网络结构

Fig. 19　Rainfall prediction CNN architecture

由于每一小时的累计降雨量对应 6 个雷达回波强度图像,所以需对每个雷达回波强度图像做如下处理:归一化的数据经过输入层后,传递到第一层卷积层,通过卷积操作和激活函数的处理后再输出到池化层,池化层下采样处理后再输出到下一个卷积层;继续通过卷积操作和激活函数的处理后再输出到下一个池化层,池化层下采样处理后的结果拉伸为一维数据后再通过第一个全连接层,经激活函数处理后得到输出数据.

6 个二维数组通过上述操作后,经由合并处理后再通过第二个全连接层,并经激活函数处理,通过输出层得到一个预测值.

本文利用 Python 与 Tensorflow 框架编写代码,将初始权重值和偏置项限制在标准正态分布的 2 倍标准差之内,期望误差、学习速率分别设置为 0.7、0.000 5,同时选择 Adam 方法为梯度下降方法,Relu 函数为激活函数.此外,第一卷积层设置的卷积核尺寸为 2×2,个数为 20 个,步长为 1;池化层采用的是最大化采样,核尺寸为 2×2,步长为 2.第二卷积层的

卷积核尺寸设置为 2×2,个数为 20 个,步长为 1,其池化层同样通过最大化采样,相应的核尺寸为 2×2,步长为 2.此外,所有全连接层节点个数都是 125.

据此构建卷积神经网络降雨量预测模型得到图 20—22.从图 20 可以看出,随着学习次数的增加,测试数据和训练数据的均方误差都是先降低然后趋于平稳,同样表明模型相当理想.而从图 21 可以看出,降雨量真实值与预测值形成的散点基本分布在 $y=x$ 周围,最终降雨量真实值与预测值的均方误差为 0.79,表明模型拟合效果较好.结合散点图和残差图,与随机森林模型和 BP 神经网络模型相比,卷积神经网络模型的预测值与真实值更加接近,且对较大的降雨强度拟合较好.

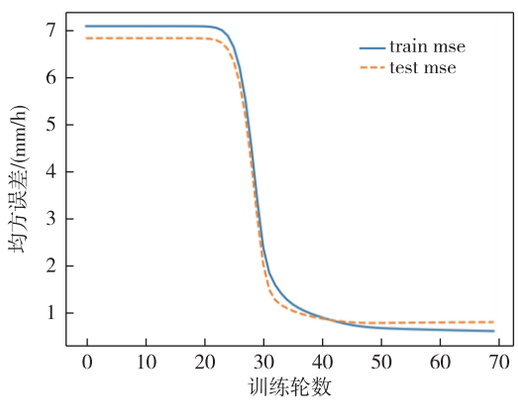

图 20　迭代过程中训练集和测试集均方误差的变化

Fig. 20　MSE variation for CNN training set and test set during iteration

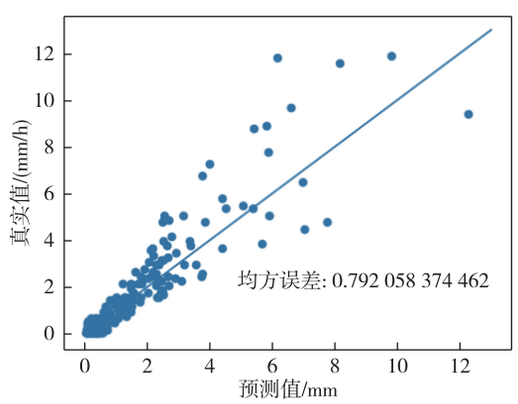

图 21　降雨量真实值与预测值的散点图

Fig. 21　Scatter of actual rainfall and CNN predicted values

4　模型预测对比

本文基于国内外学者的研究,分别选取了随机森林、BP 神经网络和卷积神经网络降雨量预测模型进行建模分析.这 3 种模型各有优缺点:随机森林方法简单高效,但对某些有特定噪声的数据进行建模时可能会出现过度拟合;BP 神经网络相对随机森林预测效果较好,但可解释性差且参数数量庞大;卷积神经网络在回归预测方面的应用较少,但它能够极大地减少复杂模型参数的数量,能够更好地挖掘特征变量之间的关系,但也有解释性差的缺点.在上述建模分析的基础上,我们利用上述的 3 个模型分别对测试集数据进行预测,预测结果如表 1 所示,而部分日期降雨实际情况和预测情况如图 23—25 所示.

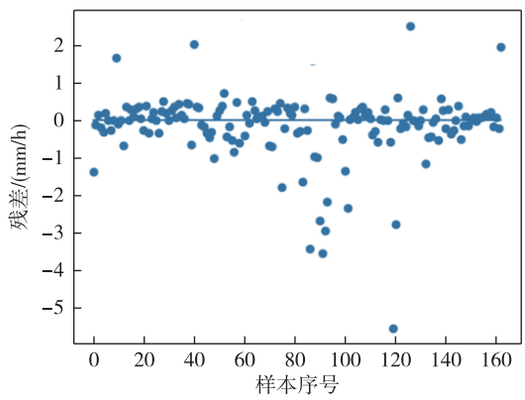

图 22　残差图

Fig. 22　Residual distribution of CNN prediction

在表 1 中,从日均方误差可以看出,随机森林降雨量预测模型和 BP 神经网络降雨量预测模型的日均方误差波动较大,而卷积神经网络降雨量预测模型的日均方误差波动相对较小.从图 23—25 可以更直观地发现,3 个降雨量预测模型对于降雨量的趋势拟合得都较好,其中卷积神经网络降雨量预测模型的预测精度最高,BP 神经网络降雨量预测模型次之,随机森林降雨量预测模型相对最差.

不过上述 3 个模型对较大降雨量的预测都不是特别好,其中随机森林模型最容易低估降雨量,BP 神经网络次之,卷积神经网络相对好点.主要原因有以下几个:一是数据的质量,本文中的数据是回波强度雷达拼图,是由多个雷达站的雷达图拼接而成,而不同雷达站的仰角不同,这就会使得回波强度的数据是在不同高度上进行拼接的,所以对降雨量的预测也是有影响的;二是样本的大小,本文的训练数据量为 2 181 个,远远达不到大样本的要求,而神经网络需要有大样本的支撑;三是模型的应用,卷积神经网络模型在回归预测方面应用很少,相关研究还不成熟.

表 1　降雨量预测均方误差比较

Table 1　MSE comparison of rainfall prediction

日期	均方误差		
	随机森林	BP 神经网络	卷积神经网络
2016-08-05	0. 005	0. 001	0. 00
2016-08-06	0. 012	0. 011	0. 02
2016-08-08	0. 010	0. 237	0. 00
2016-08-09	5. 599	5. 283	1. 05
2016-08-12	0. 003	0. 003	0. 01
2016-08-22	0. 265	5. 673	0. 01
2016-08-23	0. 004	0. 023	0. 58
2016-08-26	0. 112	0. 004	0. 11
2016-08-27	0. 007	0. 178	0. 01
2016-09-04	0. 133	0. 066	0. 09
2016-09-05	0. 080	0. 022	0. 07
2016-09-06	0. 067	0. 096	0. 08
2016-09-07	0. 047	0. 080	0. 03
2016-09-10	0. 286	0. 014	0. 41
2016-09-11	0. 240	0. 543	0. 27
2016-09-13	0. 806	0. 347	0. 51
2016-09-14	1. 824	0. 711	1. 15
2016-09-15	2. 786	2. 316	2. 13
2016-09-16	0. 752	0. 214	0. 70
2016-09-19	0. 045	0. 038	0. 02
2016-09-22	0. 044	0. 046	0. 00
2016-09-27	0. 039	0. 398	0. 37
2016-09-28	3. 721	6. 573	1. 87
2016-09-29	0. 133	0. 152	0. 12
2016-09-30	0. 729	0. 192	0. 20
2016-10-01	0. 018	0. 003	0. 01
2016-10-02	2. 414	0. 145	0. 09
2016-10-03	0. 039	0. 058	0. 02
2016-10-06	0. 250	0. 001	0. 03
总体	1. 277	1. 163	0. 79

5　结论与政策建议

降雨作为气象的重要组成部分,时刻影响着人们的生活,尤其对农业有着重要的影响,因此,准确地预测降雨量具有重要的科学意义和现实意义.

本文在国内外学者研究的基础上,对逐 10 min 的雷达回波强度数据以及气象站的逐小时降雨量数据,分别研究了随机森林、BP 神经网络和卷积神经网络在雷达预测降水量中的应用,从数据分析可以得出以下结论:

1)3 个降雨量预测模型对于降雨量的趋势拟合得都较好,其中 BP 神经网络和卷积神经网络降雨量预测模型的预测精度都要高于随机森林降雨量预测

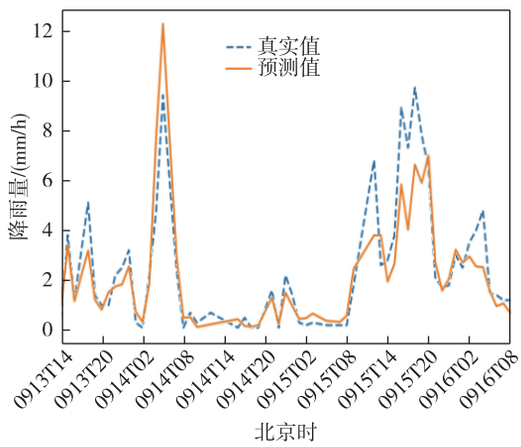

图 23　随机森林预测值与降雨量真实值的折线图

Fig. 23　Random forest predicted values and actual rainfall values

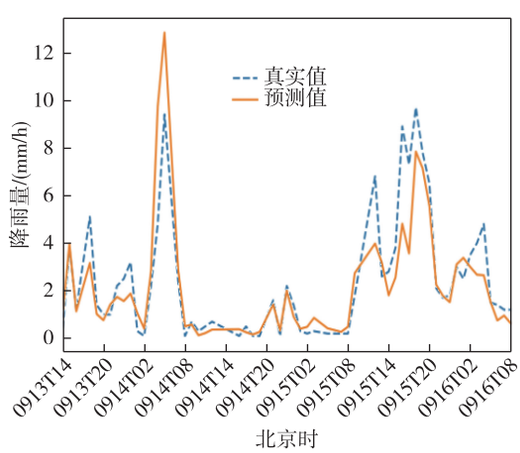

图 24　BP 神经网络预测值与降雨量真实值的折线图

Fig. 24　BPNN predicted values and actual rainfall values

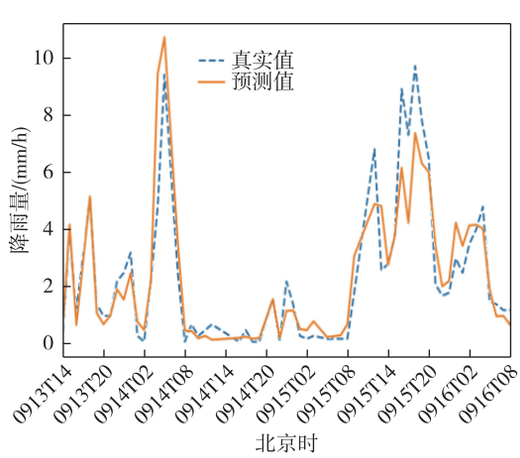

图 25　卷积神经网络预测值与降雨量真实值的折线图

Fig. 25　CNN predicted values and actual rainfall values

模型,卷积神经网络降雨量预测模型相对 BP 神经网络降雨量预测模型效果要更好.

2)相比于随机森林降雨量预测模型,卷积神经网络和 BP 神经网络降雨量预测模型每小时累计降水量预测值与实测值较为吻合,即使降水强度较大时,估测值与实测值的误差也相对较小,而随机森林降雨量预测模型对累计降雨量大于 5 mm/h 时会出现明显的低估.

3)神经网络很适合大数据的机器学习,在数据量足够多的情况下要优于随机森林.卷积神经网络由于存在局部感知,相对 BP 神经网络更能提取空间信息,因此对具有很强空间关系的降雨量预测较好.神经网络所特有的非线性特性很适合气象数据研究,但其结构和参数的选取十分关键,若选取不合理,模型的预测功能就会大打折扣.

参考文献

References

[1] Zhang G Q,Patuwo E B,Hu M Y.Forecasting with artificial neural networks:the state of the art[J].International Journal of Forecasting,1998,14(1):35-62

[2] Pomerleau D A.Efficient training of artificial neural networks for autonomous navigation [J]. Neural Computation,1991,3(1):88-97

[3] Baik J J,Hwang H S.Tropical cyclone intensity prediction using regression method and neural network[J].Journal of the Meteorological Society of Japan Ser II,1998,76(5):711-717

[4] Guhathakurta P.Long lead monsoon rainfall prediction for meteorological sub-divisions of India using deterministic artificial neural network model[J].Meteorology and Atmospheric Physics,2008,101(1/2):93-108

[5] 徐晓岭,王蓉华,吴慧玲,等.降水量的统计拟合[J].数理统计与管理,2010,29(6):961-969
XU Xiaoling, WANG Ronghua, WU Huiling, et al. Statistical fitting of precipitation[J].Journal of Applied Statistics and Management,2010,29(6):961-969

[6] 崔玫意,张玉虎,陈秋华.Box-Cox 正态分布及其在降雨极值分析中的应用[J].数理统计与管理,2017,36(1):8-17
CUI Meiyi, ZHANG Yuhu, CHEN Qiuhua. Box-cox normal distribution and its application in rainfall extreme value[J].Journal of Applied Statistics and Management, 2017,36(1):8-17

[7] Sulaiman J, Wahab S. Heavy rainfall forecasting model using artificial neural network for flood prone area[M]// Kim K J, Kim H, Baek N. IT Convergence and Security 2017. Springer,2018:68-76

[8] 张培昌,杜秉玉,戴铁丕.雷达气象学[M].北京:气象出版社,2012
ZHANG Peichang,DU Bingyu,DAI Tiepi.Radar meteorology[M].Beijing:China Meteorological Press,2012

[9] 陈程.卷积神经网络在气象短临预报的研究与应用[D].广州:华南理工大学,2018
CHEN Cheng. Application of convolution neural network in weathe rnowcasting[D].Guangzhou:South China University of Technology,2018

[10] 吴辰文,梁靖涵,王伟,等.基于递归特征消除方法的随机森林算法[J].统计与决策,2017(21):60-63
WU Chenwen, LIANG Jinghan, WANG Wei, et al. Random forest algorithm based on recursive feature elimination[J].Statistics & Decision,2017(21):60-63

[11] 李航.统计学习方法[M].北京:清华大学出版社,2012:100-103
LI Hang. Statistical learning methods [M]. Beijing: Tsinghua University Press,2012:100-103

[12] 李扬,张长,朱建平.融合统计思想的大数据算法[J].统计研究,2018,35(7):125-128
LI Yang,ZHANG Chang,ZHU Jianping.Statistical algorithms for big data [J]. Statistical Research, 2018, 35(7):125-128

[13] 王保贤,刘毅.基于灰色 BP 神经网络模型的人力资源需求预测方法[J].统计与决策,2018(16):181-184
WANG Baoxian,LIU Yi.Prediction method of human resource demand based on grey BP neural network model [J].Statistics & Decision,2018(16):181-184

[14] 周明非,汪西莉,王磊,等.高分辨卫星图像卷积神经网络分类模型[J].中国图象图形学报,2017,22(7):996-1007
ZHOU Mingfei,WANG Xili,WANG Lei,et al.Convolutional neural network models for high spatial resolution satellite imagery classification[J].Journal of Image and Graphics,2017,22(7):996-1007

[15] Hubel D H,Wiesel T N.Receptive fields,binocular interaction and functional architecture in the cat's visual cortex[J].Journal of Physiology,1962,160(1):106-154

[16] Fukushima K, Miyake S. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position [J]. Pattern Recognition, 1982, 15 (6): 455-469

[17] Lecun Y,Boser B,Denker J,et al.Handwritten digit recognition with a backpropogation network[J].World of Computer Science & Information Technology Journal, 1990,2:299-304

[18] 胡悦.基于卷积神经网络的股票市场择时模型:以上证综指为例[J].金融经济,2018(4):71-74
HU Yue.A stock market timing model based on convolutional neural network:take the Shanghai Composite Index as an example[J].Finance Economy,2018(4):71-74

Rainfall modeling and prediction by radar echo data based on machine learning

CHEN Xiaoping[1] CHEN Yiwang[1] SHI Jianhua[2]
1 College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117
2 School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000

Abstract The rainfall is modeled and predicted based on the radar echo intensity data during January to October of 2016 in Zhejiang province, and the prediction results are compared between random forest method, BP neural network model, and convolutional neural network (CNN) model. The results show that the random forest model is relatively low in accuracy, and is easy to underestimate large rainfall intensity. The BP neural network and the CNN method perform better than random forest method, especially the convolutional neural network model. Compared with the other two machine learning methods, the CNN is better in prediction accuracy and large rainfall intensity fitting.

Key words rainfall; BP neural network (BPNN); convolutional neural network (CNN); random forest