基于大数据分析的气象观测数据质量控制算法研究

尤嘉铖
（中国民用航空华东地区空中交通管理局，上海 200335）

摘要：目前，常用的气象数据质量控制算法均是依据统计学原理对数据的变化趋势进行分析，针对其算力无法对大数据进行处理以及实现质量控制的问题。文中将BP神经网络模型与粒子滤波算法结合，有效增强了神经网络的综合计算性能，将该算法应用于海量气象数据处理，可以显著减少外在因素对数据精确度的影响。实验结果表明，所提模型的预测数据误差平均值和异常数据监测误差平均值在对比算法中最小，证明该模型可以准确地监测气象数据，同时能够对异常气象数据进行修补拟合，进而实现对气象数据的质量控制。

关键词：气象数据；数据质量控制；BP神经网络；粒子滤波；数据关联度；自动气象站
中图分类号：TP391 **文献标识码：**A **文章编号：**1674-6236(2022)11-0103-05
DOI：10.14022/j.issn1674-6236.2022.11.022

Research on meteorological observation data quality control algorithm based on big data analysis

YOU Jiacheng
(*CAAC East China Regional Administration*, *Shanghai* 200335, *China*)

Abstract: At present, the commonly used meteorological data quality control algorithms are based on the statistical principle to analyze the change trend of the data, aiming at the problem that its computational power can not process the big data and realize the quality control. In this paper, the BP neural network model is combined with particle filter algorithm, which effectively enhances the comprehensive computing performance of neural network. The application of this algorithm to massive meteorological data processing can significantly reduce the impact of external factors on data accuracy. The experimental results show that the average error of prediction data and the average error of abnormal data monitoring of the proposed model are the smallest in the comparison algorithm, which proves that the model can accurately monitor the meteorological data, and can repair and fit the abnormal meteorological data, so as to realize the quality control of the meteorological data.

Keywords: meteorological data; data quality control; BP neural network; particle filtering; data correlation degree; automatic weather station

气象数据在民用领域和军用领域有着举足轻重的地位，是从事一切气象研究的基础。气象数据的质量对气象业务的发展有重要的影响[1-2]，同时气象数据的质量对于气象决策类业务以及预报类业务也有至关重要的作用，例如气象预报和气象预测的准确性。目前，我国已建立了多个地面自动气象观测站[3]，地面自动气象观测数据起到了重要的作用，这些数据是科研、气象服务工作的一手资料。因此快速、准确地对大量气象数据进行分析处理是有必要的，这也对气象数据的质量控制提出了更高的要求。

气象数据的一个显著特性即混沌性，主要原因

收稿日期：2021-03-04 **稿件编号：**202103047
基金项目：国家自然科学基金委员会与中国民用航空局联合基金(U1511328)
作者简介：尤嘉铖(1991—)，男，江苏靖江人，工程师。研究方向：气象自动观测系统。

是天气变化受各种因素影响较大，数据本身会存在着一定的误差。而数据稳定性通常受环境以及不同观测时间的影响，同时，天气的瞬时变化也会干扰测试数据，因此对气象数据进行质量控制也是一个需要解决的问题。目前常见的气象数据质量控制算法均是根据历史数据的统计分析结果来估算未来的变化趋势及走向。这种质量控制算法的门限值较宽，所以无法检测数据值小的波动，也无法处理当前海量的气象数据。该文在大数据的背景下将BP神经网络与粒子滤波算法进行结合，以对气象数据进行质量控制[4]。

1 数据质量控制模型

1.1 BP神经网络模型

BP神经网络[5-6]包含3层结构，分别为输入层、中间层及输出层，其本质是多层前馈网络。其中输入层作为数据的输入接口，然后输入接口的元胞将数据传送到中间层，中间层一般用来做数据的处理，最后将数据传送到输出层。BP神经网络结构图如图1所示。

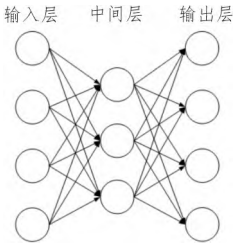

图1 BP神经网络结构图

图1中BP神经网络输入数据为向量，假设该向量为：

$$X\in R^{n},X=(x_0,x_1,\cdots x_{n-1})^{\mathrm{T}} \tag{1}$$

假设中间层中有n个神经元，则中间层的输出为：

$$x'\in R^{n_1},x'=(x_0,x_1,\cdots,x_{n-1})^{\mathrm{T}} \tag{2}$$

若输出层有m个神经元，则输出为：

$$Y\in R^{m},Y=(y_0,y_1,\cdots y_{m-1})^{\mathrm{T}} \tag{3}$$

假定输入层至中间层的权重因子为w_{ij}，阈值因子为θ_j，输出层至中间层的权重因子为w_{jk}，阈值因子为θ_k，则各个神经元输出应为：

$$\begin{cases} x_j^{'}=f\left(\sum\limits_{i=0}^{n-1}w_{ij}x_i-\theta_j\right),j=0,1,\cdots n-1 \\ y_k=f\left(\sum\limits_{j=0}^{n_1-1}w_{jk}x_j^{'}-\theta_k\right),k=0,1,\cdots m-1 \end{cases} \tag{4}$$

其中，$f(\ )$为激活函数。激活函数共有两种，分别是单极性激活函数和双极性激活函数，如式(5)、(6)所示：

$$f(x)=\frac{1}{1+e^{-x}} \tag{5}$$

$$f(x)=\frac{1-e^{-x}}{1+e^{-x}} \tag{6}$$

通常情况下，BP神经网络的传递分为两个过程：1)数据的正向传播；2)误差的反向传播[7-8]。正向传播即从输入层到中间层再到输出层的过程，反向传播将实际输出结果和神经网络输出结果的值进行误差比较，并反馈至神经网络中，神经网络会不断学习更新权值。BP神经网络结构简单，且预测准确度较高，因此在工程领域有着广泛的应用。

BP神经网络的关键是误差学习过程，下面对其误差学习过程进行说明。

BP神经网络的学习过程通过计算均方差的值进行实际计算，假设共有N个样本数据，与其对应的期望值分别为$e^{(1)}$、$e^{(2)}$、$e^{(3)}$、…、$e^{(N)}$，BP神经网络中单个神经元的输出值与理论数据的误差平方和为：

$$E^{(N)}=\frac{1}{2}\sum_{k=0}^{m-1}(e_k^{(N)}-y_k^{(N)})^2 \tag{7}$$

则BP神经网络总的误差平方和为：

$$E_A=\sum_{N=1}^{N}E^{(N)}=\frac{1}{2}\sum_{N=1}^{N}\sum_{k=0}^{m-1}(e_k^{(N)}-y_k^{(N)})^2 \tag{8}$$

网络权重的修正值应当按照梯度下降法去设定，修正值为：

$$\Delta w_{sp}=-\eta\frac{\partial E_A}{\partial w_{sp}} \tag{9}$$

由此可知，最终的BP神经网络算法流程如图2所示。

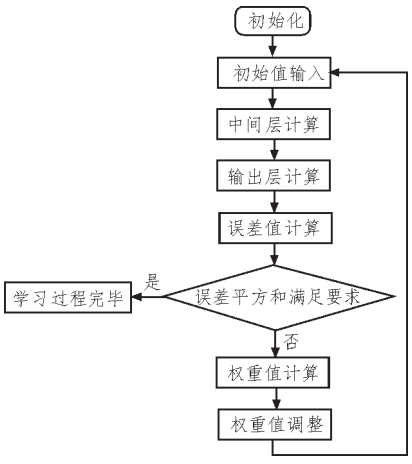

图2 BP神经网络算法流程

由以上分析可知，BP神经网络最适合于计算内部机制较为复杂的模型。由于气象数据具有混沌性的特点，因此使用BP神经网络可以对大量气象数据进行学习。但BP神经网络也有缺点，即在考虑数据相关联因素过多的情况下，系统的运算时间就会迅速变长，所求解出的数据结果的冗余度也会显著升高。

1.2　粒子滤波算法模型

由1.1节可知，BP神经网络在考虑数据相关联因素过多的情况下，系统运行速度和计算性能就会变差。因此该节引入粒子滤波算法，其算法可以对BP神经网络搜索过程中的阶梯步长进行控制，进而减少模型冗余计算，加快运算速度[9-12]。

粒子滤波算法的本质为反复迭代后得到稳定值的过程，因此粒子滤波算法的核心方程为状态方程和量测方程，如式(10)、(11)所示：

$$X_k = f_k(x_{k-1}, v_{k-1}) \tag{10}$$

$$Z_k = h_k(X_k, n_k) \tag{11}$$

式(10)为状态方程，X_k为粒子的下一个状态，f_k为状态函数，v_{k-1}为当前状态下的噪声值。式(11)为量测方程，Z_k为量测方程下一个时间的量测数据，h_k是量测系统的测量函数，n_k为下一时刻的噪声值。

粒子滤波算法的运行过程为：

1)系统初始化。模型的下一时刻设定为k，下一时刻的状态为X_k，则X_k应由$k-1$时刻的粒子状态去估算。

2)粒子采样过程。由式(10)状态方程对系统中粒子的下一时刻状态进行预测，采样的粒子为：

$$\{x_{k|k-1}(i); i = 1,2,\cdots N\} \sim P(X_K | X_{K-1}) \tag{12}$$

3)粒子权重计算过程。对粒子的预测值和粒子的实际值误差进行估算，然后计算权重，权重值为：

$$\omega_k^{(i)} = P(Z_k | X_k^{(i)}), i = 1,\cdots N \tag{13}$$

误差值越小，证明数据越准确。因此该粒子在系统中所占据的权重就越大，将实际的权重值进行归一化可得：

$$\tilde{\omega}_k^{(i)} = \frac{\omega_k^{(i)}}{\sum_{i=1}^{n} \omega_k^{(i)}}, \sum_{i=1}^{n} \tilde{\omega}_k^{(i)} = 1 \tag{14}$$

4)粒子重采样过程。对权重进行排序，舍弃掉权重小的粒子，重复过程2)。

5)系统结果输出。使用状态方程对过程4)中得到的粒子权重进行估计，估计值见式(15)所示，最终输出的值为误差值较小的粒子。

$$\hat{x}_k = \sum_{i=1}^{N} X_{k|k-1}^{(i)} \tilde{\omega}_k^{(i)} \tag{15}$$

1.3　结合粒子滤波算法的BP神经网络模型

最终的求解模型将粒子滤波和BP神经网络模型相结合，将当前状态下神经网络的权重作为粒子滤波当前系统的状态值X_k，同时结合噪声值v_k，由状态方程即可计算下一个时刻的状态X_{k+1}。将此状态权值和输入层的数据值作为激活函数$f(\ )$的变量值，最终将会求得神经网络的输出值。粒子滤波神经网络模型流程图如图3所示。

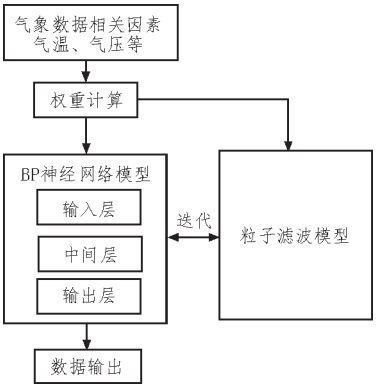

图3 算法模型流程图

2　实验分析

2.1　气象数据选取

该文气象数据选择某自动气象站点某月的气象数据，这些数据在实验前均经过测试验证，气象数据共有2 000条，其中训练集数据为1 200条，测试验证集数据为800条。表1为该次气象数据选取及环境配置的情况。

表1　数据环境配置

参数	内容
训练集/条	1 200
测试集/条	800
软件环境	Matlab R2016b，64 bit
硬件环境	I5 9400F，64 G内存

2.2　气象数据预处理

对气象数据进行质量控制前需要计算气象数据各个要素之间的关联度，分析关联度之后进行筛选，关联性较强的数据作为神经网络模型的输入[13-16]。首先对数据进行无量纲化处理，然后计算关联度，如式(16)、(17)所示：

$$x_i(k) = \frac{X_i(k)}{X_i(1)} \tag{16}$$

$$r_i=\frac{1}{N}\sum_{k=1}^{N}\xi_i(k) \tag{17}$$

其中，k为数据集合中数据的序号，r_i为序列关联度，关联度值越趋近于1表明要素关联性越强。最终计算得到的关联度如表2所示，然后将这些因素作为神经网络的权重值。

表2　气象数据关联度

参数	数值
气温	0.81
气压	0.80
湿度	0.79
风向	0.50
风速	0.92
降雨量	0.94

2.3　实验测试与结果分析

气象自动观测站站点的气象数据有多种错误，例如数据漏测、数据突变错误、数据一致性错误等。因此该文模型将对上述几种典型错误进行质量控制实验。

首先对模型的有效性进行实验测试。该文以温度数据为例进行测试，验证模型的预估值与实际值之间的曲线，然后计算方差值。同时进行对比实验，对比算法为传统统计学算法、卡尔曼滤波算法以及该文算法。实验结果如图4所示。

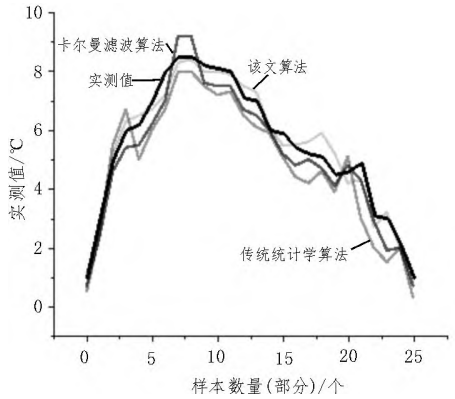

图4　模型有效性实验结果

计算得到的数据误差值，如表3所示。

表3　模型有效性实验数据误差值统计

算法	误差平均值/℃
该文算法	0.307
传统统计学算法	0.708
卡尔曼滤波算法	0.507

由曲线可以看出，文中设计的模型产生的预测曲线和实际观测值曲线接近，验证了该文算法的准确性。同时，与其他算法相比，文中算法的误差平均值最低，证明了该文算法可以对数据质量进行控制。

在数据质量控制实验中，该文选择数据跳变错误进行验证，首先将正常连续的数据插入误差值，然后再将数据输入到对比实验模型中。最终的实验结果如图5所示，拟合误差统计结果，如表4所示。

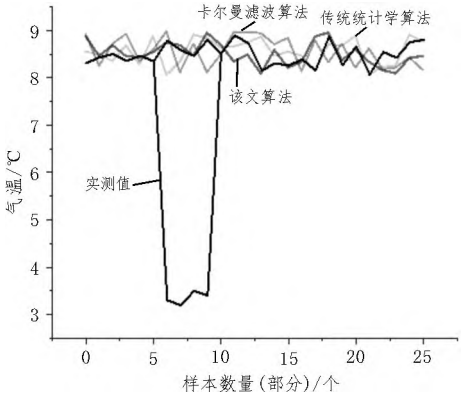

图5　数据质量控制实验结果

表4　数据质量控制实验误差统计

算法	拟合误差值/℃
该文算法	0.072 2
传统统计学算法	0.400 0
卡尔曼滤波算法	0.364 0

由图5可以直观地看出，该文提出的算法模型对数据异常处的数据拟合最为接近。因此文中算法模型可对数据跳变进行监测，同时也可以进行数据的拟合，并对气象数据进行良好的质量控制。

从数据测试结果来看，使用传统统计学算法、卡尔曼滤波算法以及该文算法的拟合数据误差为0.072 2 ℃、0.400 0 ℃、0.364 0 ℃，这说明文中模型相比其他两个传统模型对于气象数据的质量控制效果更优。

3　结束语

传统的气象数据质量控制方法无法对当前大量的气象数据进行分析以及质量控制，该文使用粒子滤波算法对传统BP神经网络模型进行优化，大幅提升了BP神经网络模型的性能。实验分析表明，文中算法在数据质量控制过程中的误差值均较小，可以有效地对气象数据进行质量控制。在未来的研究中，将进一步优化该模型的计算能力，以期在较短的时间内完成海量的数据处理，提高分析结果的实时性。

参考文献：

[1] 王力,杨福兴,曹锦飞.一种地市级气象数据共享系统的设计与实现[J].计算机技术与发展,2020,30(4):200-205.

[2] 张恩红,尹海燕,李高洁.基于Elasticsearch的气象数据检索技术研究[J].计算机技术与发展,2019,29(11):154-158.

[3] 沈瑱,王冰梅,邹超,等.无人值守地面气象观测站的设计与实现[J].信息技术与信息化,2020(5):212-216.

[4] 刘文.气象传感网中BP神经网络插值算法研究[D].南京:南京信息工程大学,2017.

[5] Dimililer K,Kiani E.Application of back propagation neural networks on maize plant detection[J].Procedia Computer Science,2017,120(7):1356-1369.

[6] 陈通,周晓辉.基于BP神经网络的深层感知器预测模型[J].计算机与数字工程,2019,47(12):2978-2981,3009.

[7] 景立森,丁志刚,郑树泉,等.基于NAG的BP神经网络的研究与改进[J].计算机应用与软件,2018,35(11):272-277.

[8] 张红玉,丁宁,徐江荣.BP神经网络激励函数改进研究[J].杭州电子科技大学学报(自然科学版),2017,37(6):62-66,90.

[9] 朱苗苗,潘伟杰,刘翔,等.基于BP神经网络代理模型的交互式遗传算法[J].计算机工程与应用,2020,56(2):146-151.

[10] 李增刚,王正彦,孙敬成.基于FPGA的手写数字BP神经网络研究与设计[J].计算机工程与应用,2020,56(17):251-257.

[11] 闫驰.基于PSO-BP神经网络的无线传感器网络定位算法[J].电子科技,2016,29(4):56-58,62.

[12] Xu Y X,Jasra A.Particle filters for inference of high-dimensional multivariate stochastic volatility models with cross-leverage effects[J].Foundations of Data Science,2019(1):201-229.

[13] 赵蕊娟.数据挖掘技术在气象预测中的应用[D].天津:天津工业大学,2017.

[14] 李飒.基于关联规则的学习行为关联度分析方法研究[J].微电子学与计算机,2018,35(6):65-68.

[15] 张岐山,郑丽君.基于灰关联分析的V-MDAV算法研究[J].计算机应用研究,2020,37(1):107-111.

[16] 牛海玲.雾霾与气象要素数据流间的关联性挖掘及应用研究[D].长春:长春工业大学,2016.

（上接第102页）

[6] 凌鹏,席文强.基于ZigBee技术高精度的输电杆塔倾斜监测预警系统[J].电子测试,2020(9):16-18,51.

[7] 徐良骥,刘悦,谌芳,等.基于GNSS-R技术的矿区复垦地土壤湿度反演方法研究[J].煤炭科学技术,2020,48(4):129-135.

[8] 章国勇,陆佳政,李波,等.电网山火同步卫星监测影像快速投影定位方法[J].高电压技术,2019,45(2):80-87.

[9] 黄亦鹏,李万彪,赵玉春,等.基于雷达与卫星的对流触发观测研究和临近预报技术进展[J].地球科学进展,2019,34(12):57-71.

[10] 葛慧磊,余翔,李永垒.基于AHP-SWOT的中国卫星导航专利技术产业化发展战略研究[J].情报杂志,2019,38(5):42-48.

[11] 秦红磊,谭滋中,丛丽,等.基于ORBCOMM卫星机会信号的定位技术[J]. 北京航空航天大学学报,2020,46(11):4-11.

[12] 李程,宋胜武,陈卫东,等.基于三维电子罗盘的边坡变形监测技术研究——以溪洛渡水电站库区岸坡为例[J].岩石力学与工程学报,2019,38(1):101-110.

[13] 杨博,王浩帆,苗峻,等.基于卫星编队的空间碎片视觉高精度导航方法[J].中国空间科学技术,2019,39(1):40-48.

[14] 马榕嵘.石墨碳纤维接地体在10 kV输电线路铁塔上的应用[J].内蒙古电力技术,2020,38(2):84-86.

[15] 葛雯斐,牛小骥,蒋郡祥,等.智能手机利用二维码路标进行定位定姿的方法研究[J].传感技术学报,2019,32(12):58-65.

[16] 葛慧磊,詹爱岚,寇冬雪.卫星导航产业技术创新态势及发展对策研究——基于专利情报多维测量[J].情报理论与实践,2020,43(3):69-74.