文章编号：2096-1618(2023)04-0392-06

基于BP神经网络气象数据质量控制方法研究

黄胃建， 杨笔锋， 卢会国， 魏春梅
(成都信息工程大学电子工程学院，四川 成都 610225)

摘要：气象观测数据的准确与否与天气、气候等预报的准确性有直接联系。气象观测数据质量控制主要是为了确保数据能具有代表性、准确性和比较性。对中国传统的质量控制算法做综述，提出存在的一些问题。在传统质量控制算法基础上，提出基于BP神经网络一致性检查的新质量控制算法，对气象观测数据实现了更精确的质量控制。

关 键 词：质量控制；一致性检查；BP神经网络
中图分类号：P468.0 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2023.04.003

0 引言

气象观测作为气象研究领域最重要和最基本的任务之一，是天气预报和气候分析数据基础和气象预报验证的重要标准。当今，气象数据的采集往往是由多个层面及多种因素决定的，因此应确保采集数据的准确性。在掌握了具备代表性、准确性和比较性的气象观测数据之后，才能获得反映实际天气现象的特征和规律[1]，以及得到较准确的天气现象的预测结果。而气象数据质量控制的意义是衡量气象数据是否具有代表性、准确性和比较性。在天气预报出现之前，认识到质量控制的重要性，质量控制的算法主要分为界限值检查和一致性检查等。

1 传统质量控制算法

受硬件的缺陷、设备的故障、局部大气扰动或自动观测站稳定性等影响，采集到的观测数据与实际数据之间存在不可避免且多样化的误差。根据气候学、大气科学、气象学和天气学等学科原理，结合站点本身的气候背景，可对上述与实测数据存在误差的观测数据进行检查，以检测出缺测和错误的数据。目前传统质量控制算法主要有：气象学界限值有效性检查，内部、时间以及空间等一致性检查[2]，如图1所示。

根据要素在气候学和物理学不可能出现的值，对观测数据进行检查称之为有效性检查，适用于所有的观测数据检查。主要分为气象学界限值有效性检查和不同台站的台站界限值检查。若未能通过有效性检查的数据则标记为错误的数据。

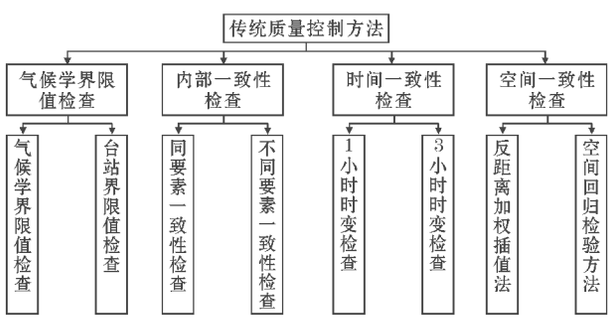

图1 传统质量控制算法框架

一致性检查主要有时间一致性检查和内部一致性检查。一致性检查适用于没有突变的观测要素。使用一致性检查可对观测数据进行更精准的质量控制，若没有通过一致性检查的数据会将质控位标记为可疑[3]。

空间一致性检查主要是利用不同气象要素在不同的地理空间所具备的特点而进行的一致性检查[4]。常用的主要方法有：反距离加权插值法、空间回归检验法[5]和Madsen-Allerupt方法[6]等。同样若没有通过空间一致性检查则将数据标记为可疑。

在传统质量控制算法中气候学界限值检查与时间一致性检查均是利用统计历史资料的特征对观测数据进行质控。内部一致性的检查是根据一定的规则对观测数据进行质控，如定时气压低于日最高气压而高于日最低气压，空气温度高于露点温度等[7]。上述检查属于界限值的质控。通过对传统质量控制算法的进一步研究发现，传统质量控制算法多依赖于从历史资料中统计的界限值。而传统方法采用的质量控制界限值区间较大，对要素的实时数据异常变化缺乏灵敏性，并且忽略了不同气象站点分布的特点与局部天气现象之间的逻辑关系。

收稿日期：2022-08-17
基金项目：国家自然科学基金资助项目(42075129)

2　基于 BP 神经网络一致性检查算法

基于上述传统质量控制算法的缺陷,充分利用气象站点分布的特点与局部天气现象之间的逻辑关系。在传统算法的基础上扩充一种基于 BP 神经网络一致性检查的算法。利用气象站点所能获得的多要素气象数据进行不同要素之间的一致性检查(图 2)。

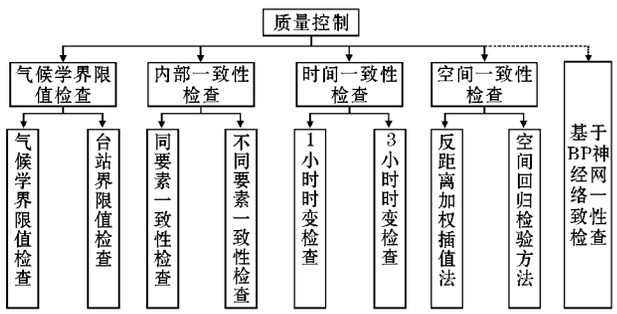

图 2　质量控制算法框架

2.1　基于 BP 神经网络内部一致性检查流程

首先对原始数据进行预处理,根据数据的质控码含义删除错误与可疑的数据,只保留正确的数据。采用最大信息系数计算不同气象要素间的相关性,然后根据计算出的最大信息系数作为衡量气象要素间的相关性。最后构建并训练 BP 神经网络模型,模型训练好之后使用验证集测试网络模型是否满足要求,如果满足要求则保存模型,反之则调整参数后继续训练,如图 3 所示。

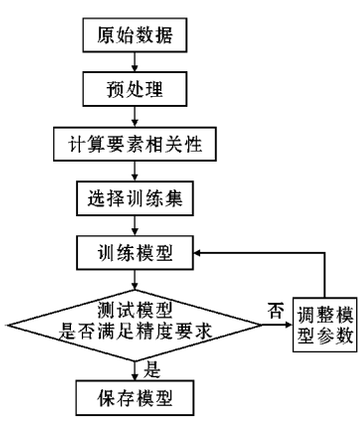

图 3　基于 BP 神经网络内部一致性检查流程

2.2　最大信息数

相关性分析是指对两个或多个存在着相关性的要素进行分析,因此可得到要素之间的相关程度。本文采用具有普适性和公平性优势的最大信息数[8](maximum information coefficient,MIC)来描述两变量间的相关性。MIC 原理为:如果两变量间存在着某种关联关系,便可使用特定规格的网格对两变量联合样本的散点图进行划分[9],利用网的边际概率密度函数和联合概率密度函数[11],最后将其归一化可得到能够检测两变量关联性的结果。MIC 的计算分为 3 个步骤:

步骤 1　对于一个给定的有限有序样本集 D,将 X 轴和 Y 轴进行划分,可获得 $x\times y$ 个网格[12],遍历不同的划分方式,可得到不同的划分方式下的互信息:

$$I(X,Y)=H(X)-H(X|Y) \tag{1}$$

式中,$H(X)$为随机变量 X 的信息熵[10],$H(X|Y)$为条件熵,即在给定 Y 的条件下 X 的信息熵。因此将式(1)按照熵的定义以及带入边缘概率分布$p(x)$,$p(y)$以及联合概率分布 $p(x,y)$展开可以得到:

$$I(X,Y)=\sum_{x,y}p(x,y)\lg\frac{p(x,y)}{p(x)p(y)} \tag{2}$$

步骤 2　将式(2)的计算结果标准化来消除划分规格对计算结果造成的影响,可得到 $x\times y$ 划分规格下的最大标准化互信息值 $m_{x,y}$($m_{x,y}\in[0,1]$):

$$m_{x,y}=\frac{\max\limits_{G}I_G}{\lg\min\{x,y\}} \tag{3}$$

式中 I_G 为某一划分方式下的未标准化的互信息,并且在计算 G 的互信息时,概率分布为落入网格中的样本量占总样本量的比率[13];

步骤 3　重复步骤 1、2,遍历不同的划分规格下的互信息值 $m_{x,y}$,则

$$\mathrm{MIC}=\max_{m_{x\times y}\in\boldsymbol{M}}m_{x\times y} \tag{4}$$

式中 $\boldsymbol{M}$ 为不同的划分规格下的互信息值 $m_{x,y}$的特征矩阵。

在实际计算中,对于任一样本数量固定的集合,其划分规格有无穷多种。因此,采用 $xy<B$,其中 B 是关于样本数量 n 的函数,按照经验通常有:$B=n^{0.6}$。随着样本数量的增加,MIC 具有如下性质:MIC=1 时,可认为两个变量间存在无噪声影响的函数关系,且无论该函数是何种形式[13];MIC=0 时,两变量相互独立[13]。

2.3　不同气象要素间相关性分析

选取自动气象站 2021 年 1 月 1 日 0 点 0 分到 12 月 31 日 23 时 59 分的分钟数据作为原始数据,包含有:2 min平均风向、风速,分钟内最大瞬时风速与其对应的风向,分钟降水量,气温,相对湿度,本站气压,草面温度,地表温度,不同深度的地温(5 cm、10 cm、15 cm、20 cm、40 cm、80 cm、160 cm、320 cm),1 min平均能见度,总辐射辐照度,反射辐射辐照度,直接辐射辐照度,散射辐射辐照度,净全辐射辐照度,共 24 个原始要素,经预处理后共 523728 条数据。选取气温、相

对湿度、本站气压以及分钟降水量作为例子,可得出MIC,如表 1 所示。

表 1 不同要素间的 MIC 值

数据	气温	相对湿度	本站气压	分钟降水量
2 min 平均风向	0.03	0.09	0.01	0
2 min 平均风速	0.06	0.12	0.02	0
分钟内最大瞬时风速的风向	0.03	0.08	0.03	0
分钟内最大瞬时风速	0.07	0.13	0.02	0
分钟降水量	0.01	0.01	0.01	0.09
气温(百叶箱)	1	0.16	0.57	0.01
相对湿度	0.16	1	0.03	0.01
本站气压	0.57	0.03	1	0.01
草面温度	0.7	0.2	0.4	0.01
地表温度(铂电阻)	0.77	0.2	0.44	0.01
5 cm 地温	0.91	0.12	0.57	0.01
10 cm 地温	0.83	0.07	0.58	0.01
15 cm 地温	0.78	0.05	0.58	0.01
20 cm 地温	0.75	0.05	0.57	0.01
40 cm 地温	0.7	0.1	0.57	0.01
80 cm 地温	0.58	0.1	0.47	0.01
160 cm 地温	0.39	0.07	0.38	0.01
320 cm 地温	0.24	0.05	0.25	0.01
1 min 平均能见度	0.18	0.24	0.08	0.01
总辐射辐照度	0.13	0.27	0.03	0.01
反射辐射辐照度	0.13	0.27	0.04	0.01
直接辐射辐照度	0.09	0.27	0.02	0
散射辐射辐照度	0.12	0.25	0.03	0
净全辐射辐照度	0.14	0.24	0.07	0.01

2.4 基于 BP 神经网络的一致性分析

1986 年由 Leema N 等[14]提出一种多层前馈型网络,即 BP(back propagation)神经网络,该网络是按照误差的逆向传播进行算法的训练。BP 神经网络的网络拓扑结构由单层的输入层、多层或单层的隐层和单层的输出层组成。每层神经网络都是由神经元构成的,并且相互连接,为全连接。它的学习规则是:为使神经网络的误差平方和最小,因此采用最速下降法,并且通过反向传播不断调整神经网络的权值和阈值[15]。如图 4 所示,BP 神经网络拥有 n 个输入层,s 个隐藏层以及 m 个输出层。

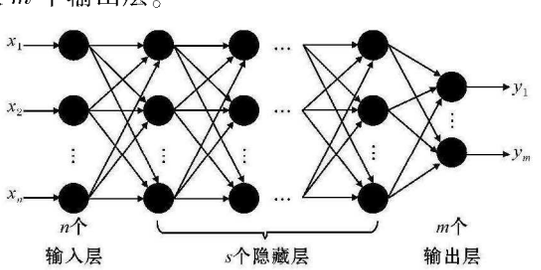

图 4 BP 神经网络结构示意图

将当前分钟气温值作为输出层(target),以上述与气温最大信息系数大于0.1的要素即相对湿度,本站气压,草面温度,地表温度,5 cm、10 cm、15 cm、20 cm、40 cm、80 cm、160 cm以及320 cm地温,1 min平均能见度,总辐射辐照度,反射辐射辐照度,散射辐射辐照度,净全辐射辐照度当前分钟的数据以及前1 min的气温值共 18 个要素作为输入层进行验证气温基于 BP 神经网络一致性分析的可行性。

根据经验公式:$H=2N+1$,其中 H 为隐藏层神经元个数,N 为输入层神经元个数。选择 37 个神经元作为隐藏层神经元个数,并选取 tansig 作为隐藏层传递函数,purelin 函数为输出层传递函数,以及训练函数 trainlm。然后使用 plotperf 函数绘制网络性能图,如图 5所示。蓝色为训练集,绿色为验证集,红色为测试集。在网络迭代了 19 次后 MSE 达到一个最小值 1.0235e-06,再利用 plotregression 函数绘制线性回归如图 6 所示。

训练集,验证集,测试集及所有的数据一起拟合的结果的相关系数均达0.9998以上。将此时训练得到的 BP 神经网络模型的参数进行保存,以便后续调用模型。

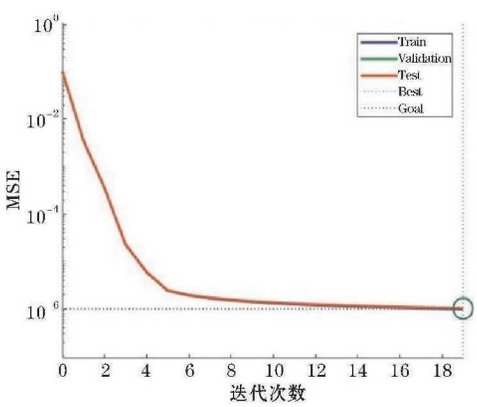

图 5 气温拟合结果 MSE

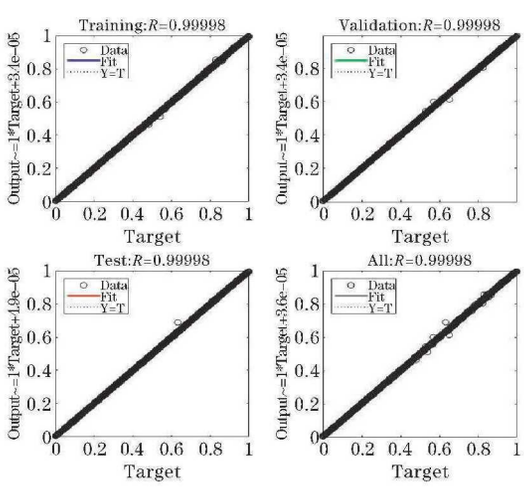

图 6 气温拟合结果线性回归

2.5　模型验证

随机选取去除缺测和漏测的 2022 年 1 月的分钟数据共 10000 条来进行 BP 神经网络的模型验证。将上述 19 个要素作为输入层,调用模型可得到预测的当前分钟气温的值,并与真实的气温数据做比较。如图 7 所示,图中蓝色代表实际数据的真实气温值,红色代表利用网络模型得出的气温预测值。可看出预测值与真实值的分布基本一致。并且决定系数 $R^2=0.9998$ 说明预测值与真实值十分接近。相对误差与绝对误差的分布如图 8、图 9 所示。可看出模型预测出的气温值与真实值的相对误差最大为0.1641,在第 6029 条的数据时得到;最大的绝对误差为0.1604,在第 5344 条的数据得到。由上述分析可以得到,利用 BP 神经网络可将不同要素内部一致性检查进一步地精确。并且根据文献[16]可知,气温采集在平均时间1 min内的准确度为0.2 ℃。因此当实测的分钟气温值与模型预测的气温值二者间的绝对误差大于 0.2 ℃时,则认为该数据是存疑的。

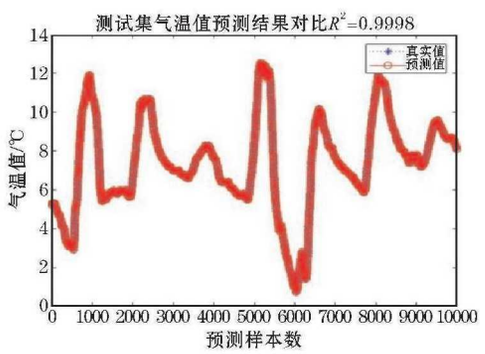

图 7　气温预测值与真实值对比

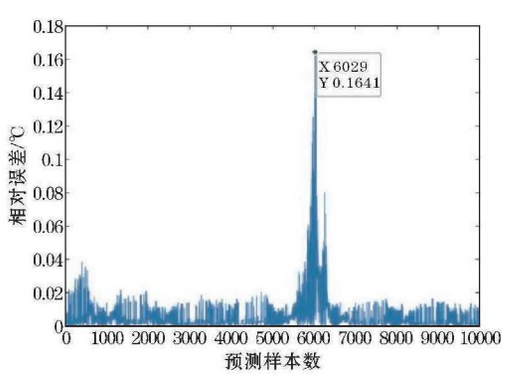

图 8　气温预测值与真实值的相对误差

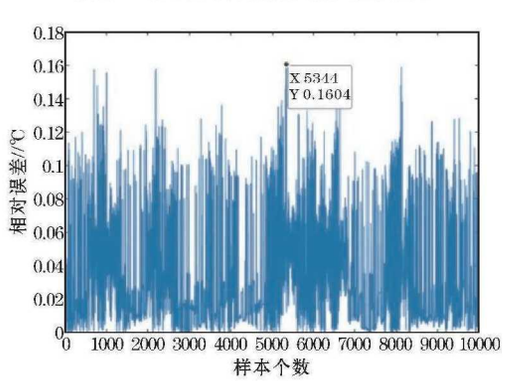

图 9　气温预测值与真实值的绝对误差

3　相对湿度基于 BP 神经网络的一致性分析

选取与相对湿度间的 MIC>0.1的要素,即2 min平均风速,分钟内最大瞬时风速,气温,5 cm、40 cm、80 cm地温,草面温度,地表温度,1 min平均能见度,净全辐射辐照度,散射辐射辐照度,总辐射辐照度,反射辐射辐照度,直接辐射辐照度共 14 个要素的当前分钟数据以及相对湿度的前 1 分钟数据作为输入层;当前相对湿度的分钟数据作为输出层。训练出的模型及误差如图 10 ~ 14 所示。根据文献[16]可知,相对湿度的采集在平均时间 1 min 内的准确度为 4%。因此当实测出来分钟相对湿度值与模型预测出的相对湿度值二者间的绝对误差大于 4% 时,则认为该数据是存疑的。

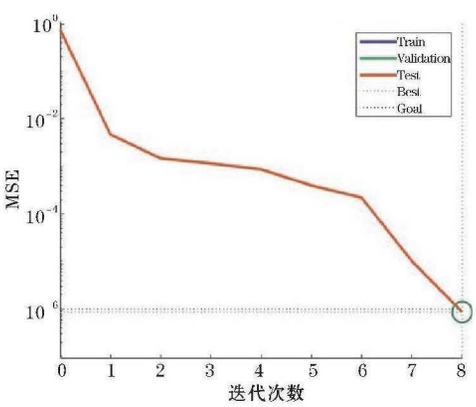

图 10　相对湿度拟合结果 MSE

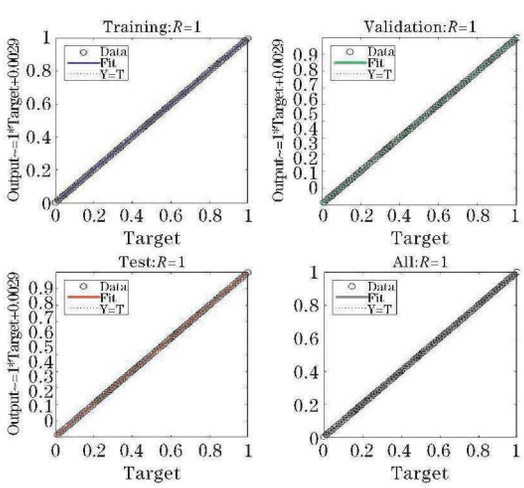

图 11　相对湿度拟合结果线性回归

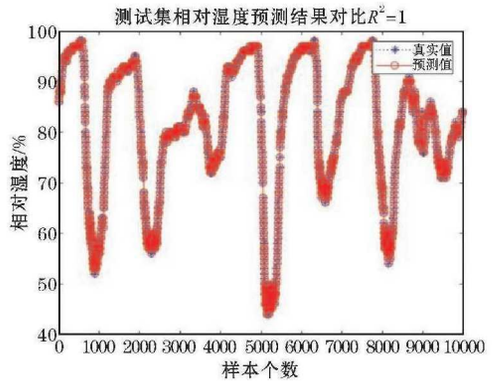

图 12 相对湿度预测值与真实值对比

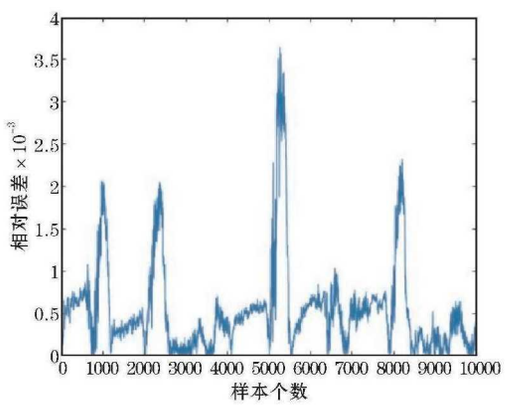

图 13 相对湿度预测值与真实值相对误差

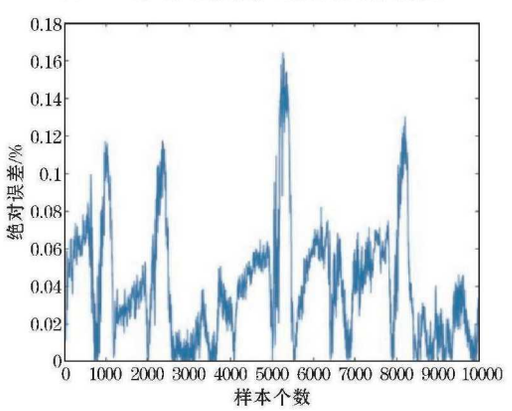

图 14 相对湿度预测值与真实值绝对误差

4 本站气压基于 BP 神经网络的一致性分析

选取与本站气压间的 MIC 数>0.1的要素,即气温,5 cm、10 cm、15 cm、20 cm、40 cm、80 cm、160 cm 以及320 cm地温,草面温度,地表温度共 11 个要素的当前分钟数据以及本站气压的前1 min的分钟数据作为输入层,本站气压当前的分钟数据作为输出层。训练出的模型及误差如图 15 ~ 19 所示。并且根据文献[16]可知,本站气压的采集在平均时间1 min内的准确度为0.3 hPa。因此当实测出来分钟本站气压值与模型预测出的本站气压值二者间的绝对误差大于0.3 hPa时,则认为该数据是存疑的。

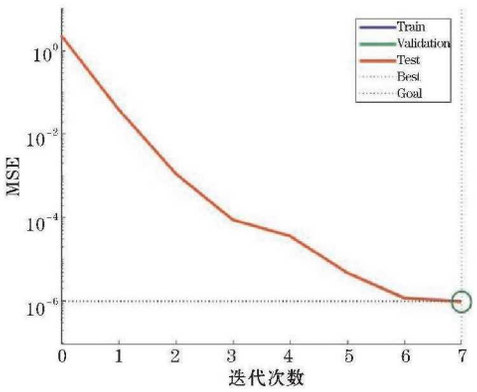

图 15 本站气压拟合结果 MSE

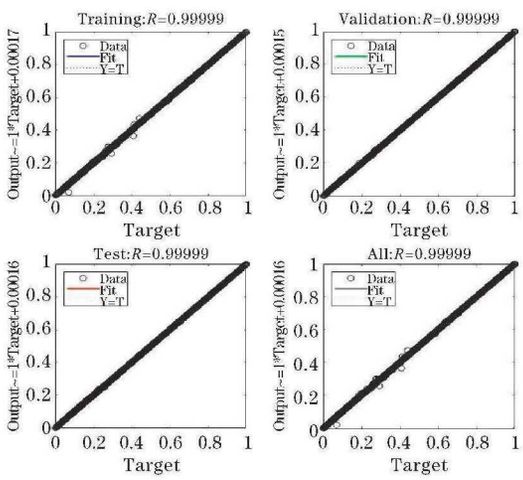

图 16 本站气压拟合结果线性回归

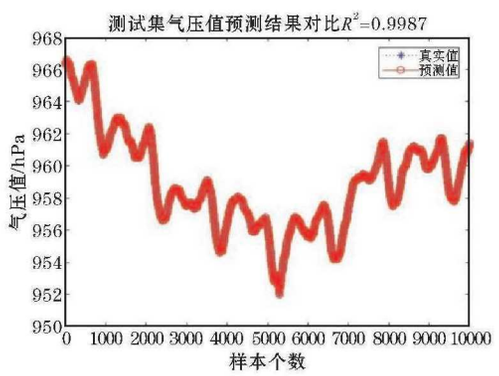

图 17 本站气压预测值与真实值对比

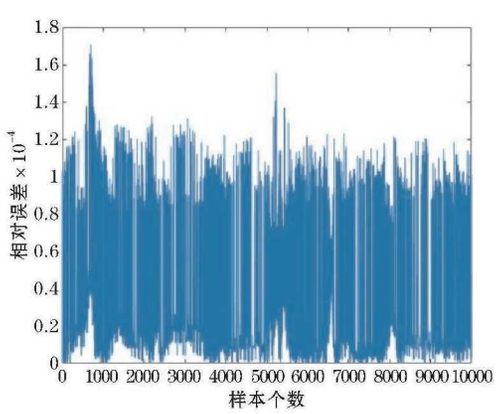

图 18 本站气压预测值与真实值相对误差

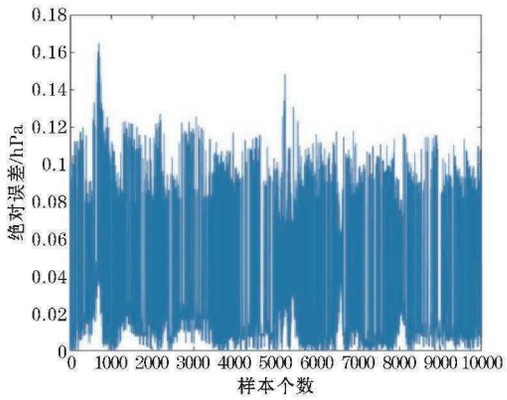

图 19　本站气压预测值与真实值绝对误差

5　结束语

介绍了一种传统的质量控制的方法,并在此基础上以气温、相对湿度、本站气压的质量控制作为示例,提出一种基于 BP 神经网络的不同要素内部一致性检查方法。与传统方法中分钟气温数据的质量控制方法相比,该方法实现了更高的精确度。如今计算机技术高速发展可为中国气象观测遇到的问题提供一种新方法,使中国的气象观测领域也能与时俱进。

参考文献:

[1] 刘晓梅. 地面自动站资料三维变分同化方法研究[D]. 南京:南京信息工程大学,2009.

[2] 李娟. 基于数据挖掘的气象观测数据质量控制算法研究[D]. 南京:南京信息工程大学,2015.

[3] 范宏飞. 自动气象站资料质量控制系统设计[J]. 吉林大学学报(信息科学版),2021,39(4):470-478.

[4] 李巍. 综合气象观测系统运行监控平台—运行监控子系统的研究与设计[D]. 北京:北京邮电大学,2012.

[5] 刘小宁,鞠晓慧,范邵华. 空间回归检验方法在气象资料质量检验中的应用[J]. 应用气象学报,2006(1):37-43.

[6] Lanzante J R. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data[J]. Internatinal Journal of Climatology,1996(16):1197-1226.

[7] 范文波. 地面气象观测数据综合质量控制方法研究与实现[D]. 南京:南京信息工程大学,2016.

[8] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large datasets[J]. Science,2011,334(6062):1518-1524.

[9] 孟燕霞. 最大信息系数算法研究[D]. 太原:太原理工大学,2019.

[10] 邵长龙,孙统风,丁世飞. 基于信息熵加权的聚类集成算法[J]. 南京大学学报(自然科学版),2021,57(2):189-196.

[11] 孟燕霞,郭禹辰,王莉. 一种基于动态均分的最大信息系数改进算法[J]. 山东大学学报(工学版),2019,49(5):105-111.

[12] Yin X R. Canonical correlation analysis based on information theory[J]. Journal of Multivariate Analysis,2004,91(2):161-176.

[13] 袁开蓉. 地面气象要素关系研究及应用[D]. 西安:西安建筑科技大学,2020.

[14] Leema N, Nehemiah H K, Kannan A. Neural network classifier optimization using Differential Evolution with Global Information and Back Propagation algorithm for clinical datasets[J]. Applied Soft Computing,2016,49:834-844.

[15] 周志华著. 机器学习[M]. 北京:清华大学出版社,2016.

[16] 中国气象局编. 地面气象观测规范[M]. 北京:气象出版社,2003.

Research on Meteorological Data Quality Control Method based on BP Neural Network

HUANG Weijian,　YANG Bifeng,　LU Huiguo,　WEI Chunmei

(College of Electronic Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: The accuracy of meteorological observation data is directly related to the accuracy of weather and climate prediction. Meteorological observation data quality control is mainly to ensure that the data can be representative, accurate and comparative. This paper summarizes the traditional quality control algorithms in China and summarizes some existing problems. Based on the traditional quality control algorithm, a new quality control algorithm based on BP neural network consistency check is proposed, which realizes more accurate quality control of meteorological observation data on the basis of the traditional quality control.

Keywords: quality control; consistency check; BP neural network