文章编号:1671-4598(2024)11-0041-07 DOI:10.16526/j.cnki.11-4762/tp.2024.11.006 中图分类号:TP391 文献标识码:A

基于卫星遥感监测极端气象预报数据异常值检测方法

李春艳
(黑龙江省牡丹江市气象局,黑龙江 牡丹江 157000)

摘要:在遥感数据采集过程中,由于传感器故障、气象条件等原因,可能会导致少量的异常点出现在采集的数据中,这些异常点可能会对极端天气预报的准确性产生负面影响;为此,需要研究一种基于卫星遥感监测极端气象预报数据异常值检测方法;基于改进K-均值聚类算法对缺失的卫星遥感监测极端气象预报数据进行插补,还原数据完整性;划分卫星遥感监测极端气象预报数据的区段,提取每个区段的裕度指标、偏斜度、频率歪度、重心频率4个特征参数,以此为输入,利用蝙蝠算法优化BP神经网络识别异常区段;计算异常区段中每个卫星遥感监测极端气象预报数据的局部离群因子,局部离群因子大于1.0数据为气象预报数据异常值,以此完成气象预报数据异常值检测;结果表明:所提方法插补误差小于±1.0,可以准确识别异常区段中的异常值,且在不同样本中的协调指数高于0.8,检测效果更好。

关键词:卫星遥感监测;极端气象;预报数据;异常区段识别;异常值检测

Detection Method for Outliers in Extreme Weather Forecast Data Based on Satellite Remote Sensing Monitoring

LI Chunyan
(Mudanjiang Meteorological Bureau of Heilongjiang Province, Mudanjiang 157000, China)

Abstract: In the process of remote sensing data collection, due to sensor failures, meteorological conditions, and other reasons, a small number of abnormal points may appear in the collected data, which may have a negative impact on the accuracy of extreme weather forecasting. Therefore, it is necessary to study a method for detecting outliers in extreme weather forecast data based on satellite remote sensing monitoring. Based on an improved K-means clustering algorithm, the missing satellite remote sensing monitoring extreme weather forecast data are interpolated to restore the data integrity. The extreme weather forecast data monitored by satellite remote sensing are divided into different sections, which extracts four characteristic parameters of margin index, skewness, frequency deviation, and center of gravity frequency for each section. Based on this input, a bat algorithm is used to optimize the BP neural network and identify abnormal sections. The local outlier factor of extreme weather forecast data monitored by each satellite remote sensing in the abnormal section is calculated. The data with a local outlier factor greater than 1.0 are considered to be abnormal values in weather forecast data, then achieving the detection of abnormal values in weather forecast data. The results show that the interpolation error of the proposed method is less than ±1.0, which can accurately identify outliers in abnormal section. Moreover, the coordination index in different samples is higher than 0.8, with a better detection effect.

Keywords: satellite remote sensing monitoring; extreme weather; forecast data; identification of abnormal sections; abnormal value detection

0 引言

气象对人们生产生活具有一定的影响,小到日常出行、大到农业种植、救灾抢险等[1-2]。气象监测最为重要的一项工作就是监测极端天气,极端天气一旦出现,往往会造成巨大的财产损失,甚至威胁民众的生命安全,如极端高温,会造成室外人员中暑,动植物死亡;如极端降雨,会造成洪水泛滥,淹没农田、庄稼,人员、动植物伤亡;如极端暴雪,会形成大范围雪灾,造成交通事故多发、冻坏农作物和牲畜。面对这种情况,极端气象预报就显得十分重要,通过气象预报能够及时进行灾害预警,降低经济损失和避免人员伤亡。随着监测技术的发展,目前气象预报数据的获取的主要手段是卫星遥感,不仅监测周期长,还实现了监测数据的可视化,监测效率也相对较高。由于天气气象预报会影响,甚至决定了很多策略的制定和实施,一旦预报数据出现异常值,则难以客观真实地反映气象的实际情况,直接导致极端气象预报不准确[3],最终造成决策失误。

面对上述问题,需要准确地检测出气象预报数据异常值。通过异常值检测可以很大程度上避免极端气象误报情况的发生。关于异常值检测,很多专家和学者都进行了相

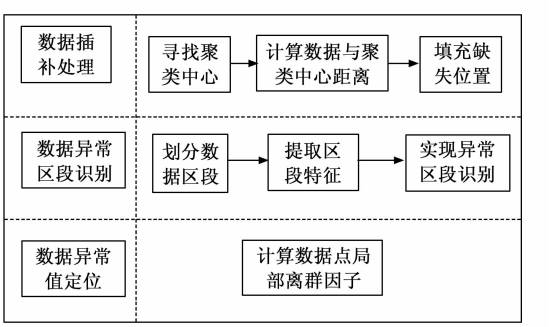
收稿日期:2023-10-18; **修回日期**:2023-12-04。
作者简介:李春艳(1983-),女,大学本科,副高。
引用格式:李春艳.基于卫星遥感监测极端气象预报数据异常值检测方法[J].计算机测量与控制,2024,32(11):41-47,55.

关研究，提出了很多检测方法。例如文献［4］以监测到的降雨数据为对象，通过 K-d tree 结构对降雨监测数据进行处理，建立了一个递进式异常筛查体系，对 K-d tree 结构上的数据进行递进式异常识别，找出存在的异常值。但是这种方法的缺点是当一组监测数据中测值中存在多个异常值时，检测效果会大大下降，容易出现漏检的现象。文献［5］提出了一种基于 SWDS-LOF 算法的异常值检测方法，该方法中利用 SARIMA 模型对监测数据进行缺失值填补处理，获取完成数据集后，通过 SWDS-LOF 算法对每个数据进行异常值判断。但该方法的缺点是检验灵敏度不高，抗干扰性不佳，容易出现误检。文献［6］提出了一种基于概率－百分位算法的异常值检测方法，该研究利用概率－百分位算法明确异常事件的时空分布特征，然后利用 GRU 构建一种识别模型，以特征为输入，实现异常识别。但这种方法对短期动态监测数据的检测能力有限。

结合前人研究经验，提出一种基于卫星遥感监测极端气象预报数据异常值检测方法。通过本研究以期提高卫星遥感监测极端气象预报数据的准确性，为极端天气预警提供可靠的参考依据。

1 卫星遥感监测极端气象预报数据异常值检测研究

为保证气象预报的准确性，去除卫星遥感监测极端气象预报数据中的异常值是十分必要的。为此，研究一种基于卫星遥感监测极端气象预报数据异常值检测方法。该方法主要分为 3 个步骤，包括卫星遥感监测极端气象预报数据插补处理、气象预报数据异常区段识别、气象预报数据异常值检测。具体框架如图 1 所示。



1 基于卫星遥感监测极端气象预报数据异常值检测框架

本次研究通过改进的 K-均值聚类算法，插补缺失的气象预报数据，划分预报数据的区段，以此提取每个区段的裕度指标、偏斜度、频率歪度、重心频率 4 个特征参数，采用蝙蝠算法优化 BP 神经网络完成异常区段的识别。通过计算异常区段中的局部离群因子，即可完成气象预报数据异常值检测。

下面针对这 3 个步骤进行具体分析。

1.1 卫星遥感监测极端气象预报数据插补处理

卫星遥感监测极端气象预报数据除了存在异常值问题外，还存在数据缺失的问题。在后期的异常值检测与分析中，需要提取预报数据的分布特征和规律，而分布特征和规律提取的前提要求是数据是完整的，否则提取到的数据特征就失去了真实性和准确性[7]。面对这种情况，在异常值检测前，需要处理好卫星遥感监测极端气象预报数据缺失的问题[8]。针对数据缺失问题，采用一种改进聚类方法进行缺失数据插补，具体过程如下：首先寻找聚类中心。聚类算法的聚类结果很容易受到初始聚类中心的影响，若是初始聚类中心选取不合适，不仅会增加算法效率，而且很容易造成过早收敛的问题，因此选择合适的聚类中心十分必要[9]。假设有一个含有 T 个预报数据对象的集合 $A=\{a_t,t=1,2,\cdots,T\}$。计算集合 A 中两两预报数据之间的距离，以两个距离最远的预报数据之间的距离作为初始聚类中心，这两个初始聚类中心记为 a_α 和 a_β，将前者作为第一个类的聚类中心，令 $a_\alpha=O_1$，将后者作为第二个类的聚类中心，令 $a_\beta=O_2$，O_1 与 O_2 之间的距离记为 d_1。计算集合 A 中剩余预报数据与 O_1、O_2 这两个聚类中心的距离，以此为依据，进行二分类，分类原则如下：计算集合 A 中剩余预报数据 a_i 与 O_1 之间的距离绝对值小于 a_i 与 O_2 之间的距离绝对值时，将 a_i 划分到 O_1 所代表的类别中，否则将 a_i 划分到 O_2 所代表的类别中，将划分出来的两个类别中记为 $A(O_1)$ 和 $A(O_2)$[10]。然后计算 $A(O_1)$ 中所有预报数据与其自身聚类中心 O_1 之间的距离并按照从大到小的顺序排列距离值，选出距离值最大值，记为 $\max d_1$。对 $A(O_2)$ 重复上述过程，得出 $A(O_1)$ 中距离聚类中心 O_2 最远的距离值，记为 $\max d_2$。对比 $\max d_1$ 和 $\max d_2$，选取二者中相对较大的距离值记为 $\hat{d}$。判断 $\hat{d}$ 是否大于二分之一 O_1 与 O_2 两个聚类中心的距离值。若大于，距离值 $\max d_1$ 和 $\max d_2$ 相对较大对应预报数据为第 3 个聚类中心，记为 O_3。重新分配气象预报数据集合 A 中剩余数据给 O_1、O_2、O_3，将划分 3 个类别，重复上述步骤，选出第 4 个聚类中心。继续重复上述过程，直至无法找到符合条件的新的聚类中心，完成聚类中心的寻找[11]。最终得到 K 个初始聚类中心，记为 $O=\{O_1,O_2,\cdots,O_K\}$。接下来利用 K-均值聚类算法对卫星遥感监测极端气象预报数据进行聚类。具体过程如下。

步骤 1：输入选出的初始聚类中心 $O=\{O_1,O_2,\cdots,O_K\}$。

步骤 2：计算除聚类中心外剩余卫星遥感监测极端气象预报数据与 K 个初始聚类中心之间的距离，记为 $d(a_t,O_k)$，$t=1,2,\cdots,T-k$。

步骤 3：根据 $d(a_t,O_k)$ 最近原则将 $\{a_t,t=1,2,\cdots,T-k\}$ 划分到靠近的聚类中心 O_k 中。重新计算聚类中心，即：

$$\overline{O}_k=\frac{\sum_{t=1}^{T_k}a_{kt}}{T_k} \tag{1}$$

式中，$\overline{O}_k$ 代表新的聚类中心；a_{kt} 代表第 k 个聚类簇中第 t 个卫星遥感监测极端气象预报数据；T_k 代表第 k 个聚类簇中气象预报数据数量。

步骤 4：判断新的类中心 $\overline{O}_k$ 是否发生改变。若改变，

回到步骤 2，否则输出聚类结果。将集合 $A=\{a_t,t=1,2,\cdots,T\}$ 划分为 K 个类别，记为 $A=\{A_1,A_2,\cdots,A_K\}$。

聚类的目的是寻找与缺失值相似的数据，以保证填充结果的准确性，因此以缺失数据所在聚类簇为基础，实施缺失数据填补，具体过程如下：首先对每个聚类簇 $\{A_1,A_2,\cdots,A_K\}$ 进行遍历查找，找出存在缺失数据的聚类簇，记为 A_k。对聚类簇 A_k 进行拆分，拆分为两个气象预报数据子集 $A_k(1)$ 和 $A_k(2)$，假设前者为完整不存在缺失的数据集，后者为不完整存在缺失的数据集[12]。$A_k(1)$ 和 $A_k(2)$ 存在以下关系：$A_k=A_k(1)\cup A_k(2)$ 且 $A_k(1)\cap A_k(2)=\phi$。计算 $A_k(1)$ 中气象预报数据的均值，将均值预填充到 $A_k(2)$ 中数据缺失位置，然后再计算 $A_k(2)$ 中气象预报数据的均值，将该均值作为最终缺失数据的插值，填补到缺失位置处。上述插值填补更适用于气象预报数据集中只存在一个缺失数据的情况，但是当一个数据集中存在两个或者两个以上的情况时，计算 $A_k(1)$ 中气象预报数据的均值，将均值预填充到 $A_k(2)$ 中数据缺失位置后，则计算 $A_k(1)$ 中气象预报数据的均值以及均值预填充位置前后的各两个数据的均值，该均值作为最终缺失数据的插值[13]。当存在连续两个缺失值时，则先同样将 $A_k(1)$ 均值作为预插值，然后再计算 $A_k(2)$ 中气象预报数据的均值，将该均值随机插入一个缺失位置处，最后再计算另一个缺失位置前后的各两个数据的均值作为插值。当存在两个以上连续缺失值时，则重复上述插值程序，直至所有缺失位置都填充完毕。如果缺失数据与均值数据相差较大，则意味着存在一些其他特征与该缺失数据相关。在这种情况下，可以尝试使用其他特征值进行预测，并采用机器学习算法建立缺失值与其他特征之间的关系，完成插值填补。

1.2　气象预报数据异常区段识别

通过卫星遥感监测可以获取大量的气象数据，但这些数据中可能存在缺失或者异常值，通过插补处理技术可以对缺失或者异常数据进行填补和修正，提高区段识别统计值的准确性，进一步提高数据的质量和可靠性，为气象预报提供更准确、可靠的数据基础。卫星遥感监测极端气象预报数据时往往是实时、连续监测的，因此获取到的数据量一般都十分巨大，若是集中进行异常值检测，工作难度大大提高，其中大量数据也会对异常值检测造成干扰[14]。针对上述问题，本章节研究在异常值检测前，先对卫星遥感监测极端气象预报数据时间序列进行区段划分并从中识别出异常区段，在最后的异常值检测中只需要针对异常区段进行检测即可，这样不仅能够缩小检测范围，大大减轻检测工作量，还能提高检测的准确性[15]。本章节研究主要分为 3 个部分的工作，即气象预报数据区段划分、气象预报数据区段特征提取以及异常区段识别。下面进行具体分析。

1）气象预报数据区段划分：

气象预报数据区段划分，也叫数据分割或者数据离散。气象预报数据本质上可以看作一个时间序列，具有时间特征。基于这种特性，采用启发式分割算法对气象预报数据区段划分。具体过程如下。

步骤 1：输入待分割的完整卫星遥感监测极端气象预报数据，记为 $A=\{a_t,t=1,2,\cdots,T\}$。

步骤 2：以 t 时刻的气象预报数据为界点，将数据集 A 分为前后两部分。

步骤 3：计算这两部分数据的平均值，分别记为 $\bar{a}_{前}$ 和 $\bar{a}_{后}$。

步骤 4：利用统计值 B 来量化表示 $\bar{a}_{前}$ 和 $\bar{a}_{后}$ 二者之间的差异。量化公式如下：

$$B=\bar{O}_k\left|\frac{\bar{a}_{前}-\bar{a}_{后}}{\Delta b}\right|=\bar{O}_k\left|\frac{\bar{a}_{前}-\bar{a}_{后}}{\left(\frac{c_{前}^2-c_{后}^2}{n_{前}-n_{后}-2}\right)^{0.5}\cdot\left(\frac{1}{n_{前}}+\frac{1}{n_{后}}\right)^{0.5}}\right| \tag{2}$$

式中，B 代表 $\bar{a}_{前}$ 和 $\bar{a}_{后}$ 二者之间的差异值；Δb 代表合并偏差；$c_{前}$、$c_{后}$ 代表 A 前后两部分的标准偏差；$n_{前}$、$n_{后}$ 代表 A 前后两部分的气象预报数据的个数

步骤 5：重复上述过程，计算每个气象预报数据的差异值 B 并从中找出差异最大值 $B_{\max}$。

步骤 6：计算 $B_{\max}$ 的统计显著性，即在随机过程中取到 B 值$\leqslant B_{\max}$ 的概率，记为 $\varphi(B_{\max})$。

步骤 7：设定一个临界值 λ，判断 $\varphi(B_{\max})$ 是否大于等于临界值 λ。若是，基于 λ 将 A 分割成两个均值有一定差异的子集合并进行下一步；若否，则不对气象预报数据集 A 进行分割。

步骤 8：对分割后的两个子集合分别重复上述分割步骤，直至所有的子集合都不能继续分割为止或者子集合的数据长度小于最小分割尺度[16]为止。

经过上述过程，将卫星遥感监测极端气象预报数据 A 分割成 N 个区段，记为 $\{D_1,D_2,\cdots,D_N\}$。

2）气象预报数据区段特征提取：

针对每个气象预报数据区段 $\{D_1,D_2,\cdots,D_N\}$，从中提取数据特征，为后续异常区段识别提供依据。正常情况下，气象预报数据呈现有规律的正态分布，存在异常值的区段，这一分布特征将发生异变，因此通过提取每个区段的特征向量，就能找出异常值所在的区段[17]。特征分为两种类型：一是时域特征，二是频域特征。为保证异常区段识别的准确性，本研究从两类特征中各提取两个参数。前者为裕度指标和偏斜度，后者为频率歪度和重心频率。

$$\text{裕度指标：}E_1=B\frac{a_{\max}(D_i)}{\frac{\sum_{t=1}^{T(D)}|a_t(D_i)|^2}{T(D_i)}} \tag{3}$$

$$\text{偏斜度：}E_2=\frac{B\sum_{t=1}^{T(D_i)}\left(\frac{a_t(D_i)-\bar{a}(D_i)}{R(D_i)}\right)^3}{T(D_i)} \tag{4}$$

式中，E_1 代表裕度指标；$a_{\max}(D_i)$ 代表区段 D_i 中气象预报数据最大值；$a_t(D_i)$ 代表区段 D_i 中第 t 个气象预报数据；$T(D_i)$ 代表区段 D_i 中气象预报数据长度；$R(D_i)$ 代表区段 D_i 中气象预报数据的标准方差；E_2 代表偏斜度。

频域特征参数在计算前，需要利用傅里叶变换将气象预报数据从时域转换到频域，转换后的频域傅里叶变换将气象预报数据表示为 $g(m)$，$m=1,2,\cdots,M$。

$$\text{重心频率：}E_3=\frac{\sum_{m=1}^{M}g(m)h(m)}{\sum_{m=1}^{M}g(m)} \tag{5}$$

$$\text{频率歪度：}E_4=\frac{\sum_{m=1}^{M}[g(m)-H]^3g(m)}{m\left(\sqrt{G_1}\right)^3} \tag{6}$$

式中，E_3 代表重心频率；$g(m)$ 代表气象预报数据的频谱序列；M 为谱线数，$h(m)$ 为第 m 条谱线的频率值；H 代表均方频率；E_4 代表频率歪度。

以上 4 种卫星遥感监测极端气象预报数据的特征参数量纲都不相同，需要利用最大—最小值方法进行统一，才能统一用于下述的异常区段识别。最大—最小值方法定义是利用计算出来的特征参数值与最小特征参数值做差，然后将差值除以最大、最小特征参数值的差值。统一量纲后，特征参数的取值会全部映射到 0～1 区间内。

3）异常区段识别：

即从 $\{D_1,D_2,\cdots,D_N\}$ 中找出存在以上的区段。在这里需要用到的识别方法为基于改进 BP 神经网络的识别模型。BP 神经网络的泛化能力强、容错性高，因此应用范围广泛，尤其在异常识别领域，更是一种常用的手段，但是该算法在训练过程中每次误差估计都要进行大量权重和阈值的调整，一旦选取不合理会导致算法早熟，进而得不到趋近于实际结果的识别结果，导致识别结果不准确[18]。面对这种情况，本研究中引入蝙蝠算法对 BP 神经网络的权重和阈值进行优化，具体过程如下：

首先设置蝙蝠算法的初始参数，然后初始化种群，将蝙蝠算法中个体的位置分量与 BP 神经网络中的权值和阈值一一对应，种群设为 $X=\{x_1,x_2,\cdots,x_N\}$。在权值和阈值的搜索空间（取值范围）内更新蝙蝠的频率、速度及位置。构建适应度函数，计算每个蝙蝠个体的适应度函数值。这里的适应度函数为每个权值和阈值设置方案下，气象预报数据训练样本输入到 BP 神经网络中，BP 神经网络得出的结果与训练样本对应的实际输出之间的均方误差函数，即

$$f(x_i)=\frac{\sum_{j=1}^{m}(I_j-\hat{I}_j)^2}{m}(E_1+E_2+E_3+E_4) \tag{7}$$

式中，$f(x_i)$ 代表第 i 个权值和阈值设置方案（蝙蝠个体）的适应度函数值；I_j、$\hat{I}_j$ 代表第 j 个训练样本的 BP 神经网络输出结果与真实结果；m 代表训练样本数量。

选出适应度函数值最小值对应的个体作为当前全局最优解，记为 $x_i(p)$，p 为迭代次数。生成一个随机数 r_1，比较随机数 r_1 与蝙蝠个体的脉冲发射频度之间的大小。若前者大于后者，则对 $x_i(p)$ 进行位置更新；若前后小于后者，则对除了 $x_i(p)$ 之外其余所有蝙蝠个体进行位置更新。生成一个随机数 r_2，判断 r_2 是否小于蝙蝠的响度且该蝙蝠适应度值是否小于当前迭代变换中所寻得的最优解的适应度值？若是，接受这个新解，更新蝙蝠个体的响度和脉冲发射速率并计算当前种群适应度值，否则重复上述过程，直至达到预设迭代次数。找出种群中适应度函数值最小值对应的蝙蝠个体，该蝙蝠个体最优位置信息对应的权值和阈值就是求取的最优权值[19]和阈值。

在获取最优权值和阈值后，利用训练样本对优化后的 BP 神经网络进行训练，使其拥有异常识别的能力。基于 BP 神经网络的识别模型如图 2 所示。

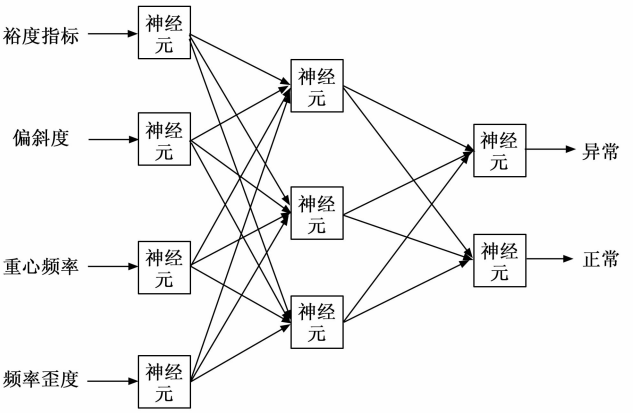


图 2 基于 BP 神经网络的识别模型

从图 2 可以看出，识别模型分为三层，即输入层、隐含层和输出层。以训练样本的输入数据的裕度指标、偏斜度、频率歪度、重心频率的标准化特征参数为输入，以发生概率为输出。训练样本的真实输出分为两类，即异常与正常，实际输出值均对应为 1，前者通过（1，0）表示，后者通过（0，1）表示。识别模型每两层之间的输入与输出表示如下：

$$\text{输入：}J_j=f(x_i)\sum_{i=1}^{M}w_{ij}E_i \tag{8}$$

$$\text{输出：}I_j=f(J_j-\varepsilon_j) \tag{9}$$

式中，I_j 代表第 j 个神经元的输入；w_{ij} 代表连接权值；E_i 代表第 i 个特征参数标准化值；M 代表神经元数量；ε_j 代表连接；$f(\cdot)$ 代表神经元激励函数；F_j 代表阈值。

计算 I_j 与训练样本真实结果 $\hat{I}_j$ 之间的均方误差并将误差沿神经网络三层结构反方向进行传播，直至均方误差的精度满足设定的精度要求。最后输入需要识别的气象预报数据区段特征值，即可得到关于该区段异常与正常两个类别的概率值，将其中较大对应的类别作为识别结果。

1.3 基于局部离群因子计算的气象预报数据异常值定位

异常区段识别是在时间序列的气象预报数据中寻找明显不同于其他数据段的部分，这些异常部分可能因为多种原因（例如，气象条件突然变化等）产生。通过识别这些异常区段，可以进一步对异常数据进行处理和修正，从而直接提高气象预报数据点局部可达密度准确性。

在经过上述章节 1.2 过程的处理和运算，从众多气象预报数据区段中找出了存在异常值的区段，但并不能准确

定位出区段中哪个值为异常值，因此还需要对识别出来的异常区段进行进一步的异常值检测。在这里用到的一种算法为局部离群因子算法。异常值必然有别于正常气象预报数据的分布特征，也就是脱离大部分的分布规律，因此可以将异常值看作有别于其他数据的离群数据[20]，示意图如图 3 所示。

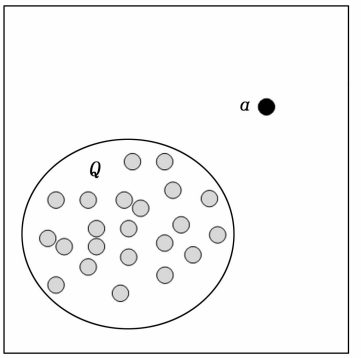


图 3　局部离群因子算法示例图

从图 3 中，可以看出点 a 与 Q 中任意一点的距离小于 Q 中任何一个数据点与其他数据点的距离，由此可以将点 s 看作了离群点。

每个气象预报数据点的局部离群因子计算过程如下。

步骤 1：输入气象预报数据异常区段，区段可以描述为 $D=\{a_t,t=1,2,\cdots,L\}$。L 代表异常区段内数据的长度。

步骤 2：计算异常区段 D 中两个数据点之间距离，记为 $d(a_i,a_j)$。

步骤 3：计算气象预报数据点 a_i 的 y －距离，记为 $y-\mathrm{dis}[a_i]$。

步骤 4：找出以距离 y 为半径的圆内关于气象预报数据点 a_i 的近邻点，这些近邻点组成邻域集合，记为：

$$Y_{y-\mathrm{dis}[a_i]}=I_j\{a_j\in(D-a_i),d(a_i,a_j)\leqslant y-\mathrm{dis}[a_i]\} \tag{10}$$

式中，$Y_{y-\mathrm{dis}[a_i]}$ 代表邻域集合。邻域集合中的气象预报数据点记为 q_j 。

步骤 5：计算气象预报数据点 a_i 与邻域集合 $Y_{y-\mathrm{dis}[a_i]}$ 内其他气象预报数据点 q_j 之间的可达距离。计算公式如下：

$$d_y(a_i,q_j)=Y_{y-\mathrm{dis}[a_i]}\max[d_y(q_j),d(a_i,q_j)] \tag{11}$$

式中，$d_y(a_i,q_j)$ 代表可达距离；$d_y(q_j)$ 代表气象预报数据点 a_i 邻域集合内点 q_j 的第 k 距离；。

步骤 6：按照下述公式计算气象预报数据点 a_i 的局部第 y 局部可达密度。计算公式如下：

$$\rho_y[a_i]=\frac{1}{\sum\limits_{q_j\in Y_{y-\mathrm{dis}[a_i]}}d_y(a_i,q_j)}\mid Y_{y-\mathrm{dis}[q_j]}\mid \tag{12}$$

式中，$\rho_y[a_i]$ 代表气象预报数据点 a_i 的第 y 局部可达密度；$Y_{y-\mathrm{dis}[q_j]}$ 代表气象预报数据点 a_i 的邻域点 q_j 的邻域集合。

步骤 6 计算气象预报数据点 a_i 的局部离群因子，计算公式如下：

$$U[a_i]=\frac{1}{\mid Y_{y-\mathrm{dis}[q_j]}\mid}\left(\sum_{q_j\in Y_{y-\mathrm{dis}[a_i]}}\frac{\rho_y[q_j]}{\rho_y[a_i]}\right) \tag{13}$$

式中，$U[a_i]$ 代表异常区段 D 气象预报数据点 a_i 的局部离群因子；$\rho_y[q_j]$ 代表气象预报数据点 a_i 的邻域点 q_j 的局部可达密度。

将离群因子大于 1.0 的气象预报数据点作为检测出来的异常值。超过 1.0，就意味着该数据点完全脱离该区段数据的正太分布规律，因此就认为该数据点为检测出来异常值。

2　实验分析

本章节以文献［4－6］中所提到的方法为对照，来验证所研究方法在卫星遥感监测极端气象预报数据异常值检测中的应用效果，以明确方法的检测性能。实验设置蝙蝠算法的搜索速度为 0.5，搜索范围为－1～1，迭代次数为 500 次；BP 神经网络设置学习率为 0.05，以此进行下述实验。

2.1　气象预报数据样本

实验测试所用到的卫星遥感监测极端气象预报数据样本均取自于中国气象网历年监测到的数据，原始数据样本均是准确且无缺失的。气象预报数据样本情况如表 1 所示。

表 1　气象预报数据样本

类别	样本长度	数据数量
高温/℃	短期	48
	中期	120
	长期	240
连续降雨/mm	短期	48
	中期	120
	长期	240
大风/(m/s)	短期	48
	中期	120
	长期	240

2.2　缺失数据插补效果分析

从表 1 每个数据样本中分别随机抽取出 1%的数据和 2%的数据，模拟数据缺失的情况，由此设置工况 1 和工况 2。利用章节 1.1 研究，改进 K-均值聚类算法先对每个数据样本进行聚类，然后找出存在缺失值的聚类簇，最后依据其研究，通过计算均值的方法进行数据插补，且为了明确章节 1.1 插补方法的效果，计算插补值与真实缺失值之间的误差，结果如表 2 所示。

表 2　缺失数据插补结果与误差计算结果

类别	样本长度	工况 1		工况 2			
		插补数据	与真实数据误差	插补数据 1	与真实数据 1 误差	插补数据 2	与真实数据 2 误差
高温/℃	短期	38.5	0.2	40.3	0.4	39.6	0.6
	中期	39.1	－0.3	41.5	0.5	40.7	－0.5
	长期	38.7	0.2	38.6	－0.7	42.3	0.2
连续降雨/mm	短期	87.63	－0.58	64.85	－0.32	88.62	0.74
	中期	75.74	0.42	87.96	－0.47	87.12	0.55
	长期	92.35	0.63	90.23	0.85	90.34	0.48
大风/(m/s)	短期	31.5	0.7	40.3	－0.4	40.1	0.3
	中期	42.2	0.6	35.6	0.5	38.9	－0.3
	长期	41.5	0.2	37.8	－0.4	44.6	－0.7

从表 2 中可以看出，插补出来的气象预报缺失数据与真实数据之间的上下误差均小于±1.0，由此说明本研究章节 1.1 所研究的插补方法达到了缺失值有效填补的目的，证明章节 1.1 研究是有效的。

2.3 气象预报数据异常区段识别效果分析

为模拟气象预报数据异常值情况，从表 1 每个数据样本中分别随机抽取出 1%的数据、3%的数据以及 5%的数据，然后在对应位置人工插补异常值，使得表 1 中原有无异常值的气象预报数据样本转变为存在异常值的气象预报数据样本。基于章节 1.2 研究，首先针对每个测试用的气象预报数据样本进行区段分割处理，以其中一个样本作为示例，区段分割结果如图 4 所示。

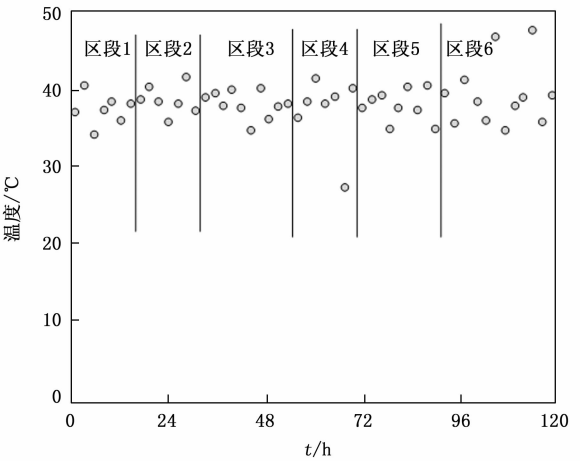


图 4 气象预报数据区段分割结果示例

从图 4 中可以看出，示例样本共划分了 6 个区段。针对每个区段，提取对应的 4 个特征参数，包括裕度指标、偏斜度、频率歪度、重心频率，然后统一每个特征参数的量纲，结果如表 3 所示。

表 3 气象预报数据区段特征参数表

区段	裕度指标	偏斜度	频率歪度	重心频率
1	0.552 6	0.176 8	0.154 3	0.054 5
2	0.218 8	0.276 8	0.356 2	0.015 4
3	0.328 2	0.321 7	0.282 0	0.049 7
4	0.585 4	0.543 2	0.262 2	0.052 6
5	0.256 8	0.278 0	0.478 3	0.075 3
6	0.682 0	0.685 4	0.397 8	0.039 8

将每个样本区段的特征参数值输入到优化 BP 神经网络当中，利用 BP 神经网络的识别能力识别出存在异常值的区段。同样以上述示例样本作为示范，气象预报数据异常区段识别结果如图 5 所示。

从图 5 中可以看出，示例样本划分出来的 6 个区段中识别出来有 2 个区段为异常概率大于正常概率，分别为区段 4 和区段 6，说明这两个区段中存在异常值，为识别出来的异常区段。

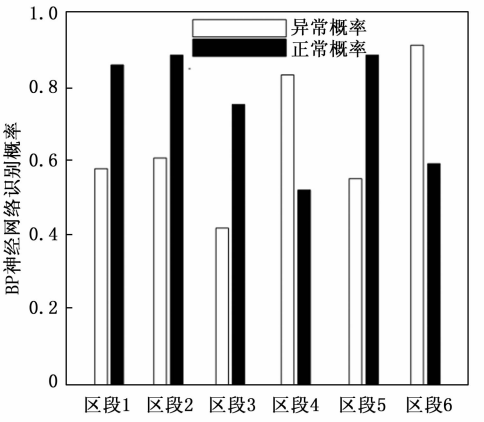


图 5 气象预报数据异常区段识别结果

2.4 异常值识别结果

基于章节 1.3 研究，计算识别出来的异常区段中每个数据的局部离群因子，作为异常值判断的依据。同样以选出的示例样本作为示范，异常区段 4 中共有 7 个气象预报数据，异常区段 6 中共有 12 个气象预报数据。这些数据的局部离群因子计算结果如表 4 所示。

表 4 异常区段中气象预报数据的局部离群因子计算结果

异常区段 4		异常区段 6	
数据序号	局部离群因子	数据序号	局部离群因子
1	0.547 8	1	0.768 2
2	0.463 3	2	0.328 2
3	1.682 3	3	0.128 3
4	0.283 3	4	0.205 4
5	0.354 0	5	1.565 2
6	0.205 4	6	0.325 5
7	0.755 6	7	0.523 3
/	/	8	0.812 3
/	/	9	1.688 2
/	/	10	0.533 3
/	/	11	0.235 3
/	/	12	0.328 8

从表 4 可以看出，示例样本异常区段 4 中序号 3 对应的气象预报数据的局部离群因子大于 1.0，异常区段 6 中序号 5 和序号 9 对应的气象预报数据的局部离群因子大于 1.0，说明这 3 个数据为检测出来的异常值。

按照上述示例样本的处理过程，对表 1 中每一个样本都进行相同的处理，得出处理结果。

2.5 方法检测性能对比分析

针对表 1 相同的实验样本，利用文献［4］～文献［6］中所提到的方法进行异常值检测，得出检测结果。根据检测结果与真实情况，统计漏检因子和误检因子，然后计算这两个因子的协调指数，以验证检测方法的检测能力。漏检因子计算公式如下：

$$Z = \frac{N_1}{N_1 + N_2} \tag{14}$$

式中，Z 代表漏检因子；N_1 代表未检测出来的异常值数量；N_2 代表检测出来的异常值数量；N_1+N_2 代表实际异常值总量。误检因子计算公式如下：

$$V=\frac{M_1+M_2}{N_1+N_2} \tag{15}$$

式中，V 代表误检因子；M_1 代表检测结果为异常数据但实际为正常数据的数据数量；M_2 代表检测结果为正常数据但实际为异常值数据的数据数量。根据漏检因子和误检因子，计算协调指数，公式如下：

$$\psi=\frac{1}{w_1Z+w_2V} \tag{16}$$

式中，ψ 代表协调指数，该指数越大，说明方法的检测能力越好；w_1 、w_2 代表漏检因子和误检因子对应的权重系数，在这里各自取值为 0.5。

文献［4－6］检测方法与所研究检测方法的协调指数对比结果如图 6 所示。

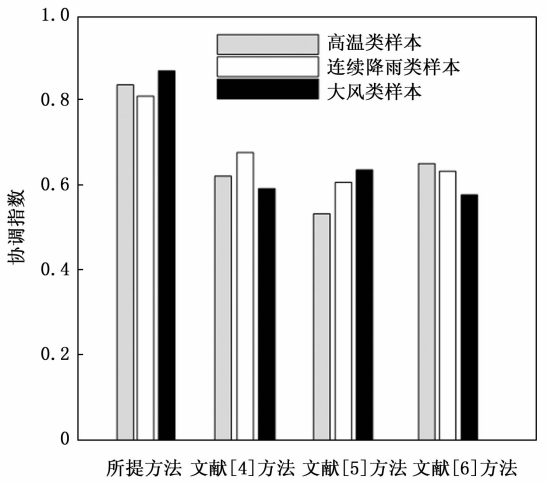


图 6　协调指数对比结果图

从图 6 中可以看出，所研究方法应用下，3 个预报数据样本的漏检因子和误检因子协调指数均要大于文献［4－6］检测方法，由此说明所研究方法的检测能力更强，更不容易出现异常值漏检和误检的情况，保证了异常值检测结果的准确性。

在上述研究的基础之上，进一步测试 4 种方法下，极端气象预报数据异常值检测时间，检测时间越短、表明方法的检测效率越高，可以在更短的时间内实现异常检测快速完成极端天气预警。检测结果如图 7 所示。

由图 7 可知，所提方法在不同样本的检测过程中，检测时间始终未超过 50 ms，检测效率最高。文献［4－6］方法虽检测时间也较短，但均高于所提方法，检测时间在 60～90 ms 之间，表明 3 种对比方法的检测效率还需进一步地提高。

3　结束语

卫星遥感是监测极端气象的主要手段，为避免监测到的气象预报数据出现错误，进行异常值检测是十分必要的。

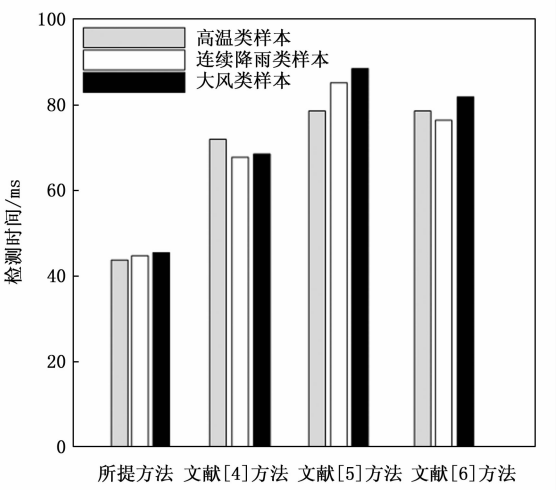


图 7　检测效率测试

为此，研究一种基于卫星遥感监测极端气象预报数据异常值检测方法。该研究主要取得 3 个成果：（1）对 K-均值聚类算法进行了改进，提高了缺失数据插补质量；（2）通过优化神经网络识别了异常区段，将异常值识别工作分散，降低了异常值识别工作量；（3）利用局部离群因子算法进一步定位出异常区段中的异常值，完成异常值检测。

参考文献：

[1] 郝　颖，车建峰，冬　雷，等. 一种基于邻域保持的数值天气预报数据降维可信度评估准则［J］. 太阳能学报，2022，43（6）：106－114.

[2] 王怡婷，陈　曦，鞠兴旺，等. 基于 GRU 网络的气象要素预测算法［J］. 计算机仿真，2021，38（7）：419－423.

[3] 李　艺，华　静，刘保双，等. 大气污染物监测数据异常值判别方法研究［J］. 环境科学学报，2022，42（12）：341－352.

[4] 田济扬，刘含影，刘荣华，等. 大规模降雨监测数据异常识别方法［J］. 中国水利水电科学研究院学报（中英文），2022，20（5）：438－448.

[5] 陆秋琴，魏　巍，黄光球. 环境监测系统中异常数据的识别和修复方法［J］. 安全与环境学报，2021，21（3）：1300－1310.

[6] 王　彤，谭索怡，吕　欣. 基于气象大数据的连续异常监测方法［J］. 中国科学院大学学报，2023，40（3）：362－370.

[7] 周笑天，陈益玲，李　芸，等. 一种基于特征曲线的自动土壤水分观测数据异常值检测方法［J］. 中国农业气象，2022，43（3）：229－239.

[8] 潘莹丽，刘　展，宋广雨. 基于 SCAD 惩罚回归的异常值检测方法［J］. 统计与决策，2022，38（4）：38－42.

[9] 李　清. 基于改进 PSO-PFCM 聚类算法的电力大数据异常检测方法［J］. 电力系统保护与控制，2021，49（18）：161－166.

[10] 何书锋，孙钿奇，王　诏，等. 基于深度学习的多波束海底地质数据异常值检测方法［J］. 计算机应用与软件，2021，38（4）：95－100.

（下转第 55 页）