文章编号:1672-6553-2025-23(9)-074-011

DOI:10.6052/1672-6553-2025-024

深度学习结合改进 DBSCAN 聚类的数据异常检测*

王典　常军†

(苏州科技大学 土木工程学院,苏州　215011)

摘要　由于结构健康监测系统采集到的数据不可避免存在异常,导致无法从中获取结构真实健康情况,故异常数据检测对结构分析及其状态评估至关重要. 为此,提出一种基于组合预测模型的多通道数据异常检测方法. 首先,将结构健康监测数据分为两段,前段只有环境引起的间歇性异常,后段包括间歇性异常以及传感器故障造成的数据异常. 其次,通过根据余弦核密度估计各数据点的局部密度自适应地选取参数半径,并对基于密度的空间聚类算法(DBSCAN)改进,进而用该改进模型剔除前段数据中的间歇性异常得到清洗数据(即没有问题的正常数据). 接着,基于多传感器间的相关性,结合卷积神经网络(CNN)的空间特征和长短期记忆网络(LSTM)的时间特征,训练清洗数据得到代表正常数据特征的数学模型. 然后,在数学模型中输入后段数据得到预测数据,并将预测数据与后段数据对比得到预测误差,采用极值理论(EVT)拟合预测误差分布并设置阈值,进而检测数据的异常状况. 最后,分析 Dowling Hall 人行天桥加速度监测数据表明,该方法能够有效提高结构健康监测异常数据的检测能力.

关键词　深度学习,　DBSCAN 算法,　数据异常检测,　组合模型,　结构健康监测

中图分类号:TU317　　**文献标志码**:A

Combining Deep Learning with Improved DBSCAN Clustering for Data Anomaly Detection*

Wang Dian　Chang Jun†

(School of Civil Engineering, Suzhou University of Science and Technology, Suzhou　215011, China)

Abstract　Due to the inevitable presence of anomalies in the data collected by the structural health monitoring system, it is impossible to obtain the true health status of the structure. Therefore, anomaly data detection is crucial for structural analysis and state evaluation. A multi-channel data anomaly detection method based on a combination prediction model is proposed. Firstly, the structural health monitoring data is divided into two sections. The first section only includes intermittent anomalies caused by the environment, while the second section includes intermittent anomalies and data anomalies caused by sensor failures. Secondly, by estimating the local density of each data point based on cosine kernel density and adaptively selecting parameter radii, and improving the density based spatial clustering of applications with noise (DBSCAN) algorithm, the improved model is used to remove intermittent anomalies in the previous data and obtain clean data (i. e., normal data without problems). Next, based on the correlation between multiple sensors, combined with the spatial features of convolutional neural networks (CNN) and the temporal features of long short term memory networks (LSTM), a mathematical model representing normal data features is trained to clean the data. Then, the predicted data is obtained by in-

2025-01-01 收到第 1 稿,2025-03-01 收到修改稿.

* 国家自然科学基金资助项目(52208189),江苏省高等学校基础科学(自然科学)(21KJB580006), National Natural Science Foundation of China (52208189), Basic Science (Natural Science) of Higher Education Institutions in Jiangsu Province (21KJB580006).

† 通信作者 E-mail:changjun21@126.com

putting the later stage data into the mathematical model, and the predicted data is compared with the later stage data to obtain the prediction error. The extreme value theory (EVT) algorithm is used to fit the distribution of the prediction error and set a threshold to detect abnormal conditions in the data. Finally, analyzing the acceleration monitoring data of Dowling Hall pedestrian overpass shows that this method can effectively improve the detection ability of abnormal data in structural health monitoring.

Key words deep learning, DBSCAN algorithm, data anomaly detection, combined model, structural health monitoring

引言

我国现有在役桥梁近 90 万座,利用桥梁健康监测系统采集的数据对桥梁结构健康状况进行评估,为桥梁运营维护部门提供维修、管理决策依据具有重要的现实意义[1]. 近年来,我国大跨桥梁安全事故发生次数逐年减少,主要得益于为这些桥梁安装的较为完善的健康监测系统[2],对桥梁在役期间进行了全周期、多方面、多种类的结构健康监测,并据此对桥梁结构进行日常养护等[3]. 然而,数据采集、传输和存储的每个环节都可能存在导致结构健康监测数据异常的因素[4],从而导致系统采集到的数据出现缺失、漂移等异常,进而无法正确判断桥梁结构损伤情况[5],故对异常数据检测势在必行[6].

随着人工智能技术的发展,深度学习技术逐渐应用于健康监测数据处理领域[7]. Smarsly 等[8]在 SHM 系统的无线传感器节点中嵌入了一个多层预训练人工神经网络(ANN),自主检测和隔离漂移和偏置故障. Fu 等[9]设计的人工神经网络能检测出峰值、漂移、偏置三种异常数据,在珍岛大桥数据应用中验证了该方法的有效性和精度. Bao 等[10]提出了一种基于计算机视觉和深度学习技术的自动异常检测深度神经网络,将一维原始时间序列信号分段转换为二维灰度图像,并对每张图像进行手工标注. 该方法在实桥中检测异常数据准确率达 87%.

在实际工程中,多种传感器相互关联,仅用单一特征检测异常状况难以奏效. 由于多变量时间序列中异常样本与正常样本在空间特征上存在明显差异,需考虑多个空间特征的相关性. 传统方法使用小部分数据应用于大数据时性能提升往往有限. 此外,研究发现,普遍用于模型训练的类不平衡数据集会对数据异常检测的准确性产生负面影响.

为解决上述问题,本文提出一种基于无监督组合预测模型的异常检测框架,可减少类不平衡数据集带来的精度下降、计算复杂等负面影响.

1　自适应参数的 DBSCAN 聚类模型

1.1　DBSCAN 算法

DBSCAN(density-based spatial clustering of applications with noise)是一种基于密度的空间聚类算法,由 Ester 等[11]在 1996 年提出. 该算法将具有足够高密度的区域划分为簇,并能在带有噪声的空间数据库中发现任意形状的聚类.

DBSCAN 算法的核心思想是:对于给定的数据集,算法首先找到核心对象,即在给定半径 ε 内包含不少于最小数量点的点. 然后,算法从这些核心对象出发,通过密度可达性关系,将紧密相连的核心对象归为同一簇. 对于那些不是核心对象的点,如果它们位于核心对象的邻域内,则也被分配到相应的簇中. 不属于任何簇的点被视为噪声.

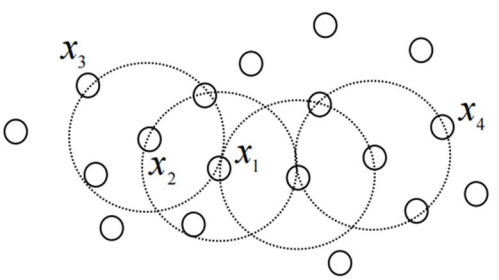


图 1　DBSCAN 核心节点邻域示意图

Fig. 1　Schematic of the neighborhood of DBSCAN core nodes

定义 1　(R −邻域半径)对$x_j \in D$,R −邻域包含样本集 D 中与x_j 之间的距离小于等于R 的样本,即:

$$N_R(x_j)=\{x_i \in \mid \mathrm{dist}(x_i,x_j) \leqslant R\} \tag{1}$$

定义 2　(核心对象)如果 x_i 的R −领域至少包含 MinPts 个样本,x_j 是一个核心对象,即:

$$\mid N_R(x_j) \mid \geqslant \mathrm{MinPts} \tag{2}$$

定义 3 (密度直达)如果 x_j 在 x_i 的 R－邻域中,并且 x_i 是核心对象,称 x_j 由 x_i 直接密度可达.

定义 4 (密度可达)对于 x_i 和 x_j,如果存在一个序列 $p_1,p_2,\cdots,p_n$,其中 $p_1=x_i,p_n=x_j$ 并且 p_{n+1} 由 p_i 直接密度可达,称 x_j 由 x_i 密度可达.

定义 5 (密度相连)对于 x_i 和 x_j,如果存在 x_k 使 x_i 和 x_j 均由 x_k 密度可达,则称 x_i 和 x_j 密度相连.

1.2 DBSCAN 算法的改进—自适应参数的 DBSCAN 算法

通过核概率密度函数进行估算得到数据集中每个样本点的核概率密度[12].研究表明,样本点密度越大,越有可能是一个聚类的中心,并且在该点邻域搜索半径(Eps)越大,同一聚类中的点可以尽可能多地聚为同一簇.当样本点的密度较小时,该点可能是簇的边缘点或离散点,设置较小的 Eps 可以降低对其他簇的影响[13].由于核密度估计不依赖于数据集分布的先验知识,因此可以利用概率密度估计值作为每个点设置 Eps 的参考,然后进行聚类.

定义 6 (核密度估计)假设独立分布 F 包含 $x_1,x_2,\cdots,x_n$ 个样本点,概率密度函数为 f,核密度估计如式(3):

$$\hat{f}_h(x)=\frac{1}{n}\sum_{i=1}^{n}K_h(x-x_i)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{3}$$

其中,h 表示带宽,$K(x)$表示核函数,同时 $K(x)$ 满足以下条件:

$$K(x)\geqslant 0,\quad \int K(x)\mathrm{d}x=1$$

$$\int xK(x)\mathrm{d}x=0,\quad \int x^2K(x)\mathrm{d}x>0 \tag{4}$$

定义 7 (核函数)核函数是从输入空间到特征空间存在映射关系的内积.

在计算样本局部密度时,不同数据集样本点局部密度差异较大,如图 2 所示.基于欧氏距离计算的局部密度难以准确刻画其分布.基于高斯核函数计算局部密度时,因考虑全局信息而增加了计算复杂度.高斯核函数曲线也直观地反映了其需要全部数据信息的特点,如图 3 所示.对比余弦核函数曲线,如图 4,它仅使用样本点局部邻域内的信息,避免了对全局数据的依赖,大幅降低了计算复杂度,使得算法能够高效运行,并能准确描述局部密度的分布.余弦核函数见公式(5).图 3 和图 4 中,横轴可代表两个样本间的距离,纵轴为核函数值.可观察到,在有效作用范围内,核函数值与间距呈负相关,当距离近时核函数值高,即相似性高的现象.

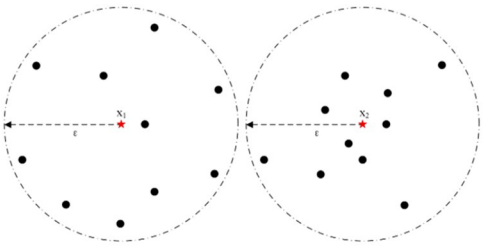


图 2 两个数据集的样本分布示意图

Fig. 2 Sample distribution of the two datasets

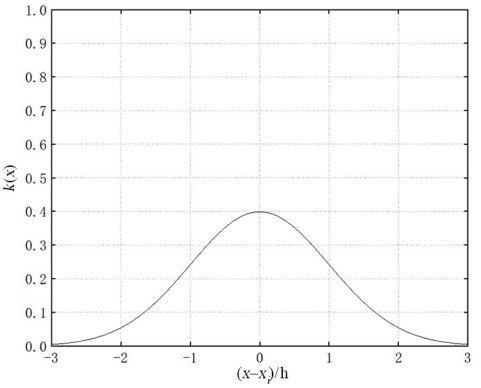


图 3 高斯核函数

Fig. 3 Gaussian kernel function

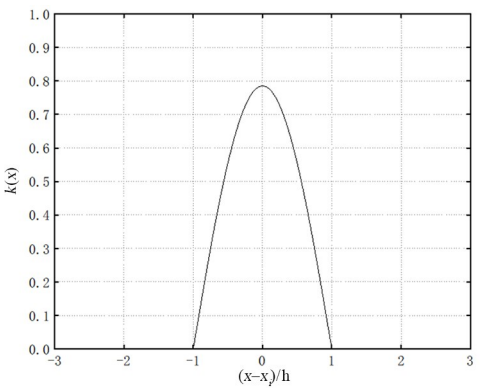


图 4 余弦核函数

Fig. 4 Cosine kernel function

$$K(x)=\begin{cases}\dfrac{\pi}{4}\cos\dfrac{\pi}{2}x, & x\in[-1,1]\\ 0, & \text{其他}\end{cases} \tag{5}$$

在 x 定义域内,两点间距离小于带宽 h 时,样本间互有影响,通过公式(5)采集局部信息定义样本局部密度,可以凸显相同截断距离内的点的位置差异;同时平衡了类簇的中心点和边界点对样本局部密度的影响.

由公式(3)～(5)得到数据每个样本点的概率

密度,概率密度的定义认为高密度区域设定较大的 ε 值捕获更大范围的结构信息;低密度区域设定较小的 ε 值以避免将噪声或远离的点错误地归入某簇.具体步骤如下:

(1)计算余弦核密度估计值.计算数据集中每个点的余弦核密度估计值.此值将用于后续确定每个点的 ε.

(2)确定 ε 的基准值和变化范围.用整体中位数作为 ε 的基准值,即 $\varepsilon_{\text{base}}$,它反映整个数据集密度状态的平均水平.

(3)为每个点定制 ε_1.对于每个点 i,基于其余弦核密度估计值 D_i 相对于最大密度 $D_{\max}$ 和最小密度 $D_{\min}$ 的位置,自适应调整 ε_i:

$$\varepsilon_i = \varepsilon_{\text{base}} \times \left(0.5 + \frac{D_i - D_{\min}}{D_{\max} - D_{\min}}\right) \tag{6}$$

ε_i 的变化范围为 $(0.5, 1.5)\varepsilon_{\text{base}}$.这样确保了 ε_i 值会根据点的密度值在制定的范围内自适应调整.

(4)确定样本数目 MinPts.基于整体数据的特性和经验规则来确定,根据文献[14]可得 MinPts=2×dim.

(5)应用于 DBSCAN.将上述步骤自适应确定的参数应用于 DBSCAN 算法.每个点将根据其特有的 ε 值来判定邻域内的点,从而实现对密度不均匀数据的更好聚类.

2　基于深度学习的数据异常检测模型

2.1　CNN-LSTM 预测模型

在结构健康监测中,CNN-LSTM 组合模型可充分利用不同传感器数据的空间关联特征和历史数据的时序演化规律,同时考虑传感器间的相互影响以及变化趋势,从而实现对目标传感器的监测预测,为健康状态评估提供可靠依据.

根据 CNN 和 LSTM 的特点,建立了基于 CNN-LSTM 的数据预测模型.模型结构图如图 5 所示,主要结构为 CNN 和 LSTM,包括输入层、一维卷积层、池化层、LSTM 隐藏层、全连接层.

基于 CNN-LSTM 进行桥梁振动数据预测的整体流程可以概括为如下关键步骤:

(1)数据预处理.剔除原始数据中间歇性异常,并归一化处理.

(2)选择滑动窗口宽度.通过不断调参选择合适的宽度.

(3)构造 CNN-LSTM 模型.选择合适的网络层、激活函数、优化器等,并调整学习率、批量大小等初始参数以提高模型性能.

(4)数据集划分与训练.将数据集划分为训练集和预测集,训练时持续更新相关参数,直到模型预测准确度满足要求.

(5)预测并计算结果.将测试集输入到训练好的模型,得到预测结果,作为数据异常预警的依据.

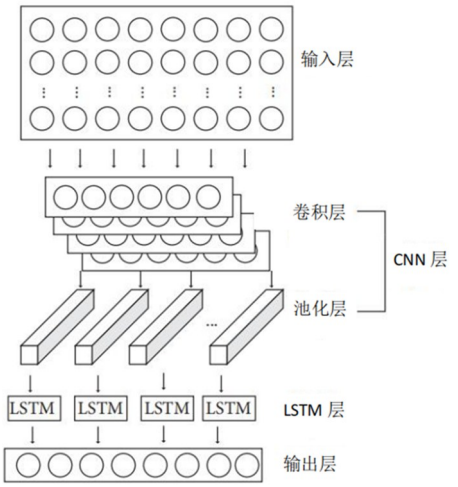


图 5　CNN-LSTM 结构图

Fig. 5　CNN-LSTM structure diagram

2.2　基于 EVT 的阈值动态控制

基于极值理论(extreme value theory, EVT)的阈值动态控制主要是在处理极端数据、高维数据分析、风险管理等领域中的一个重要应用.极值理论是统计学中用于研究和建模随机变量极端偏离其平均水平值行为的理论.在金融、气候研究、网络安全等领域,对极端事件的预测和管理至关重要.基于 EVT 的阈值动态控制策略,就是通过极值理论模型来动态确定数据处理或风险管理中的阈值,从而有效地预警和管控潜在的极端风险.与传统阈值方法相比,该方法能适应数据的非平稳性,提高异常检测的灵敏度和准确性,在结构健康监测中具有重要应用价值.

极值理论中最常用的两个分布是广义极值(generalized extreme value,GEV)分布和广义帕累托分布(generalized pareto distribution,GPD).GEV 用于建模一系列数据中的最大值(或最小值),而 GPD 用于对超过某一阈值的数据进行建模.EVT 的核心在于使用合适的模型来描述并预测极端事件的行为,特别是在超出历史数据范围的

情况下.将预测误差作为输入,来进行数据的异常检测.步骤如下:

(1)确定异常阈值.将阈值 u 设为预测误差的 95%分位数.估计 GPD 参数,计算 90%置信水平下的异常阈值 q.

(2)异常检测.对新数据,用模型预测并计算预测误差 e.若 $e>q$,发出异常警报;否则视为正常.

(3)更新模型和阈值.定期将新数据加入训练集,重新训练模型.更新预测误差、GPD 参数和异常阈值 q.

2.3 异常检测评估指标

基于解混淆矩阵,传感器异常数据检测评估指标主要有 4 种,分别为准确率(A)、精确率(P)、召回率(R)和 F_1 分数($F_{1\text{-Score}}$),如下:

$$A=\frac{T_{\mathrm{P}}+T_{\mathrm{N}}}{T_{\mathrm{P}}+F_{\mathrm{P}}+F_{\mathrm{N}}+T_{\mathrm{N}}} \tag{7}$$

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \tag{8}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \tag{9}$$

$$F_{1\text{-Score}}=\frac{2PR}{P+R} \tag{10}$$

其中,将异常数据定义为正类,正常数据定义为负类.T_{P} 为预测是异常数据实际上也是异常数据的数据量;T_{N} 为预测是正常数据实际上也是正常数据的数据量;F_{P} 为预测是异常数据实际上是正常数据的数据量;F_{N} 为预测是正常数据实际上是异常数据的数据量.

3 结构健康监测数据异常检测流程

在结构健康监测系统中,数据的采集、传输、存储往往存在导致数据发生异常的因素,进而影响后续对工程结构的分析和对结构状态的评估.基于无监督提出一种综合改进后密度聚类与深度学习的数据异常检测流程,为桥梁的准确预警提供科学依据.具体步骤如下:

(1)对安装有健康监测系统的桥梁进行数据采集;

(2)用改进后自适应参数 DBSCAN 聚类模型对正常数据进行清洗;

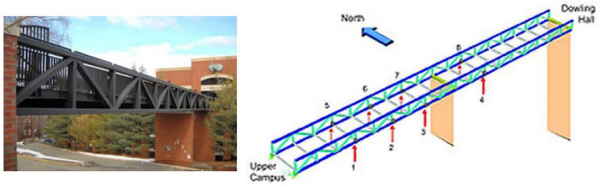
(3)把清洗后的监测数据用于 CNN-LSTM 模型训练,得到桥梁正常状态下的预测模型;

(4)对预测结果进行分析,将预测误差用于基于极值理论的阈值动态控制得到阈值,进而检测数据的异常状况;

(5)结合评价指标确定异常检测效果.

4 案例分析

实验数据采用 Dowling Hall 人行天桥健康监测(https://engineering.tufts.edu/cee/shm/research/continuous-monitoring-dowling-hall-footbridge)加速度数据训练集,其位于塔夫茨校区,用于结构健康监测相关研究与教学,8 个加速度传感器的布置如图 6 所示.该桥前三阶频率分别为 4.63 Hz、6.07 Hz 及 7.07 Hz.



(a)道林大厅人行天桥 (b)加速度传感器位置

(a)Dowling Hall Footbridge (b)Location of accelerometers

图 6 Dowling Hall 人行天桥加速度传感器布置方案

Fig. 6 Sensor placement of Dowling Hall Pedestrian Bridge

4.1 异常数据模拟及数据集制作

在实桥监测过程中,异常工况发生的概率较低,传感器收集的异常数据样本有限,且异常数据成因复杂,数据异常类型难以准确定义.为了全面评估异常检测算法的性能,使用数学模型模拟各种可能的异常情况,构建丰富的异常数据集.参考已有数学模型[15,16],对某时间序列监测数据 $x(t)=(x_1,x_2,\cdots,x_n)$,主要考虑的数据异常类型及其数学模型如表 1 所示.

表 1 数据异常数学模型

Table 1 Mathematical models for data anomalies

异常类型	数学表达
跳点	$x_i<[x_{\min}]$或 $x_i>[x_{\max}]$
增益	$x(t)=G[x_0(t)+\omega(t)]$
线性偏移	$x(t)=x_0(t)+k\cdot t+b+\omega(t)$
精度下降	$x(t)=x_0(t)+\theta(t)$
丢失	$x(t)=0$

4.1.1 数据集制作总体流程

采用 Dowling Hall 人行天桥健康监测数据构造数据集的总体流程如下:

(1)基于原始数据,采用移动窗口截取数据集,设置滑动窗口移动步长为 1024.其中,前 14 周的数据构造初始训练集,后 3 周的数据构造初始测试集;

(2)计算初始训练集的均值和方差,并利用 Z-score 方法对初始训练集和初始测试集归一化处理;

(3)对初始训练集和初始测试集的数据进行随机采样,得到训练集和测试集;

(4)对测试集,采用表 1 中的数学模型,模拟响应输入数据.

4.1.2　异常数据模拟结果

对于数据跳点,基于每个数据段的振幅来引入数据跳点噪声.在每个长度为 1024 的数据段上,随机选择 3 到 7 个数据点,并为这些点引入随机噪声,设置为振幅的 1.5 到 8 倍,如图 7 所示.

对于数据增益,主要考虑了:整体缩小、局部缩小两种增益情况.前者随机生成 0.02 到 0.3 倍的增益系数;后者随机选择 0.2 到 0.9 倍数据段长度的数据进行增益,如图 8 所示.

将线性偏移中的趋势项和数据漂移分开模拟,如图 9 所示.趋势项往往包含较长周期的干扰,因此模拟"趋势项"除了考虑表 1 中的线性偏移项,额外加入了周期性的正弦函数,具体如下:将 1024 长度的数据段随机分段(1 段到 4 段),并在每段分段上添加幅值在 0.3 到 2.5 倍振幅的线性高差(递增、递减和随机正负号组成的四种情况);在每段分段上进一步叠加随机周期、随机初始相位、随机幅值的正弦函数,其中周期范围在 0.5π 到 2.5π 之间,随机初始相位在 0 到 2π 之间,随机幅值在 0.3 到 2.5 倍的振幅之间.

数据漂移不同于数据趋势项,是监测数据的某一段均值突然增大然后迅速回归到正常数据的情况.具体步骤如下:确定漂移类型后随机选取 0.1 到 0.4 倍数据长度的数据段,并生成 0.5 到 2.5 倍的振幅的峰值点高度;随机确定局部"趋势项"的拟合函数次数,并将其叠加到原始数据上.

原始信号中已经包含有噪声,所以在以上几种异常数据模拟中均没有再次添加随机噪声.对于精度下降情况,通过在原始信号中添加随机信噪比(SNR)在−5 到 0 之间的噪声来模拟,如图 10 所示.

对于数据丢失,设置信号丢失率在 0.3 到 0.8 之间,并分为连续整段数据丢失和离散点式数据丢失两种情况模拟,如图 11 所示.

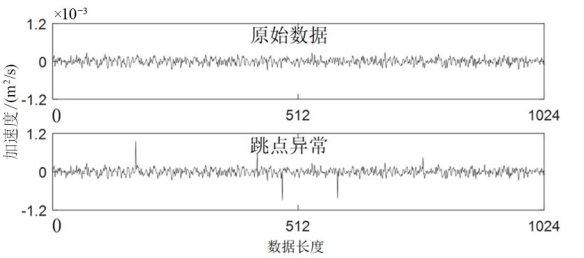


图 7　跳点异常模拟

Fig. 7　Jump point anomaly simulation

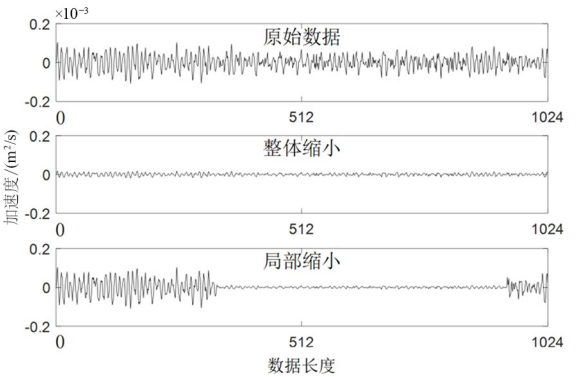


图 8　数据增益模拟

Fig. 8　Data gain simulation

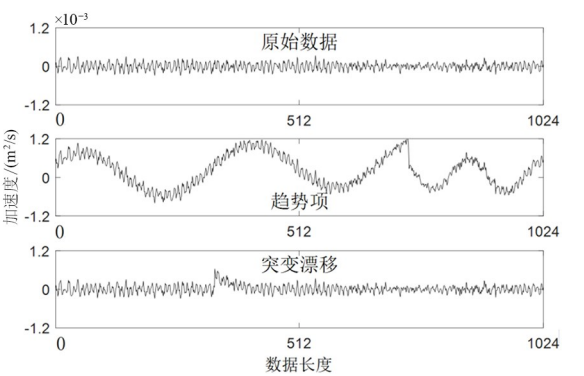


图 9　线性偏移异常模拟

Fig. 9　Linear offset anomaly simulation

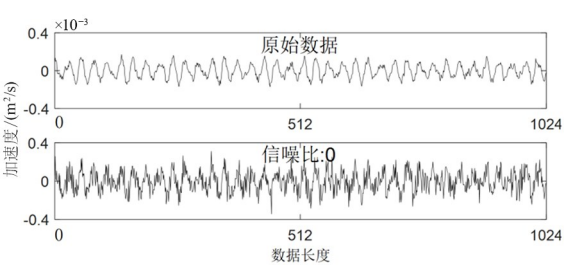


图 10　精度下降模拟

Fig. 10　Accuracy drop simulation

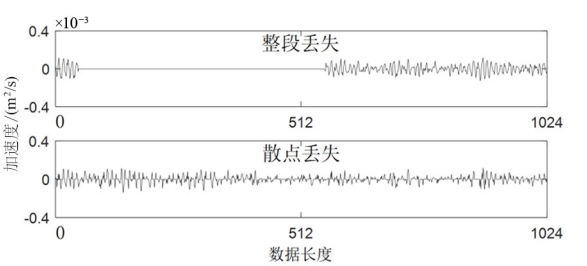


图 11 数据丢失模拟
Fig. 11 Data loss simulation

4.2 基于自适应参数 DBSCAN 算法的数据异常检测

选取该桥 2010 年 1 月 5 日 19:00～20:00 的监测数据，为了对比 DBSCAN 改进效果，分别采用改进前后的 DBSCAN 方法对异常数据进行识别，结果见图 12、图 13 和表 2. 结果表明，改进后的自适应参数 DBSCAN 算法能够更准确地识别出离群点. 从表 2 可以看出，改进的 DBSCAN 算法在异常检测的准确率、精确率和召回率均高于传统 DBSCAN 算法. 原因是传统 DBSCAN 算法中相关参数受人工选取的限制，在面对不均匀数据样本时，易把离群点错误地归入正常数据簇中，导致错检和漏检，降低了异常识别的精度. 改进后的自适应参数 DBSCAN 算法能够根据局部密度的变化自适应调整半径参数，加快聚类速度并提高识别精度.

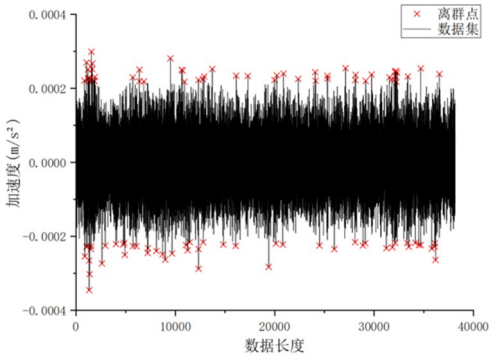


图 12 基于 DBSCAN 算法的异常检测
Fig. 12 Anomaly detection based on DBSCAN algorithm

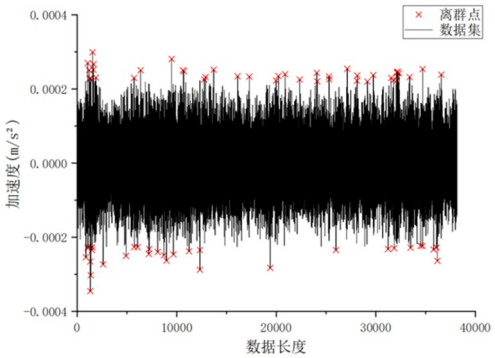


图 13 基于自适应参数 DBSCAN 算法的异常检测
Fig. 13 Anomaly detection based on adaptive parametric DBSCAN algorithm

表 2 传感器异常数据检测方法评估
Table 2 Evaluation of sensor anomaly data detection methods

评估指标	DBSCAN 算法	自适应 DBSCAN 算法
准确率(A)	0.897	0.941
精确率(P)	0.785	0.863
召回率(R)	0.744	0.859
F_1－分数($F_{1-\mathrm{Score}}$)	0.764	0.861
运算时间/s	145	139

4.3 基于组合预测模型的数据异常检测

4.3.1 数据相关性分析

桥梁健康监测系统中，多种类型的传感器分布在结构的不同测点中，通过获取桥梁结构响应数据为状态评估提供支持. 因此，分析传感器在不同测点间的关联性十分重要.

图 14 展示了数据集中传感器 s1 和 s2 的部分加速度数据随时间变化情况. 可看出，安装在同一纵梁的相邻传感器响应数据在峰值处存在部分偏差，但总体变化趋势基本一致. 为验证不同测点响应之间的相关性，计算了测点 s1～s8 加速度监测数据的相关系数，系数越大，相关性越强.

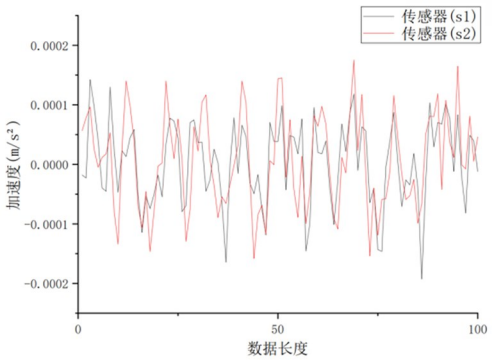


图 14 相邻传感器加速度数据随时间变化图
Fig. 14 Plot of acceleration data versus time between neighboring sensors

使用式(11)计算灰色关联系数：

$$\zeta_i(k)=\frac{\Delta(\min)+\rho\Delta(\max)}{|x_0(k)-x_i(k)|+\rho\Delta(\max)} \tag{11}$$

ρ 一般取值为 0.5，其中 $\Delta(\min)$ 和 $\Delta(\max)$ 分别是其他传感器数据与待检测传感器绝对差值的最小值和最大值.

对 i 个时刻的关联系数求和，再取其均值，可得最终的灰色关联度，即：

$$R_i=\sum_{n=1}^{k}\zeta_i(n) \tag{12}$$

选用灰色关联度作为评价数据相关性的评价指标，尽管皮尔逊相关系数可反映两个测点的相关性，但对于时滞数据可能产生不准确结果. 时滞数据是指时程曲线形态相似，但各时刻数据并不对

应,可能由传感器所测结构响应存在时间延迟所致.相比之下,灰色关联度对存在一定时滞性的时间序列具有更好的鲁棒性,故应用灰色关联度更具有适用性.具体的相关系数如表 3 所示.

表 3　相关系数表

Table 3　Correlation coefficient table

传感器	s1	s2	s3	s4	s5	s6	s7	s8
s1	1.000	0.812	0.786	0.752	0.793	0.749	0.731	0.715
s2	0.812	1.000	0.827	0.805	0.814	0.838	0.761	0.686
s3	0.786	0.827	1.000	0.774	0.708	0.721	0.740	0.733
s4	0.752	0.805	0.774	1.000	0.653	0.689	0.727	0.806
s5	0.793	0.814	0.708	0.653	1.000	0.836	0.809	0.728
s6	0.749	0.838	0.721	0.689	0.836	1.000	0.857	0.763
s7	0.731	0.761	0.740	0.727	0.809	0.857	1.000	0.748
s8	0.715	0.686	0.733	0.806	0.728	0.763	0.748	1.000

相关系数绝对值越大代表相关性越强;反之则相关性越弱.当两个应变测点监测数据的相关系数大于 0.700 时,说明它们具有强相关性.由表 3 可知,传感器 s1 与其余传感器间的相关系数均大于 0.700,保持强相关性.因此,可将加速度传感器 s2～s8 的输出数据作为输入,共同参与对加速度传感器 s1 的预测.

4.3.2　训练过程

本文基于 Pytorch 框架完成模型的搭建和训练,硬件配置 CPU 为 Intel(R) i5-8500K、RAM 为 32 GB、显卡为 NVIDIA GTX 1060.采用 Adam 算法更新网络参数,设置批尺寸为 64,设置学习率为 0.0001,设置损失函数为平均绝对误差,设置迭代次数为 100 个循环(epoch).

对 CNN-LSTM 神经网络预测模型的超参数进行合理设置,可提高模型的预测准确性.常见的模型参数设置方法有试验法、交叉验证法、网格搜索法、遗传算法等.对 CNN-LSTM 模型的超参数选择需采用分层优化策略:

(1)人工预设参数.针对桥梁监测数据的时间序列属性,采用一维 CNN(滤波器 32)提取时序数据局部特征,单层 LSTM(隐藏单元 64)建模时序依赖,避免深层网络对低频振动信号的过拟合.

(2)自动优化参数.通过网格搜索联合 K 折交叉验证($K=5$)对可调参数进行寻优,覆盖学习率(1×10^{-4}～1×10^{-2})、批处理量(32～256)、卷积滤波器数量(32～128)等.

经网格搜索和 K 折交叉验证以及人为调整后,确定的网络模型超参数选择如表 4 所示.其中对其余所提模型均进行了统一的超参数优化流程,包括训练集划分比例、评估指标均保持一致等.模型训练过程如图 15 所示,训练收敛且没有出现过拟合现象.在训练初始阶段,训练损失和测试损失均呈现明显的下降趋势.在训练 30 个循环之后网络性能的提升逐渐放缓,并在第 96 个循环得到最优的测试结果.根据分布直方图将预测误差可视化,为模型性能提供了重要定量和定性分析依据.通过直观观察误差的分布情况,可以评估模型是否具有良好的预测能力,并对其在不同数据集上的适用性进行进一步的优化和调整.如图 16 所示,CNN-LSTM 模型的预测误差接近正态分布,具有较好的预测性能.

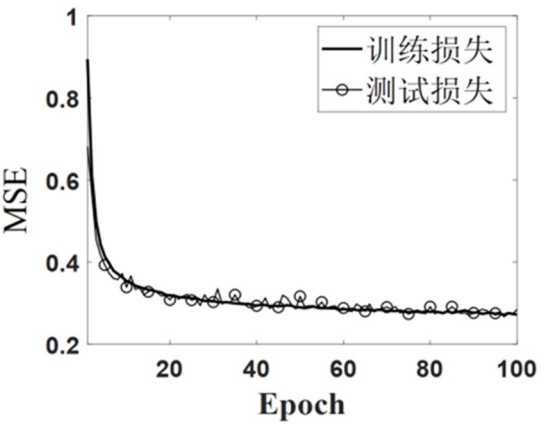


图 15　训练过程

Fig. 15　Training process

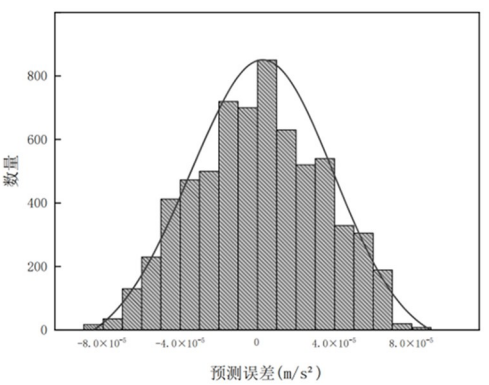


图 16　预测误差分布直方图

Fig. 16　Histogram of prediction error distribution

表 4　网络结构参数

Table 4　Network architecture parameters

参数	选择
卷积层滤波器	32
卷积层核大小	1

卷积层激活函数	Tanh
池化层大小	1
池化层激活函数	Relu

续表 4

参数	选择
LSTM 层的隐藏单元数	64
LSTM 层激活函数	Tanh
批处理大小	64
学习率	0.001
优化器	Adam
损失函数	平均绝对误差

4.3.3 数据预测对比结果分析

为验证自适应 DBSCAN-CNN-LSTM 组合模型的适用性，将自适应 DBSCAN-CNN-LSTM 组合模型预测结果与 LSTM 模型、CNN-LSTM 模型、RNN 模型在相同的训练数据集和实验数据集下进行预测结果对比实验. 为了更加清晰地显示模型预测效果，选取部分测试数据集进行训练效果展示，如图 17 所示.

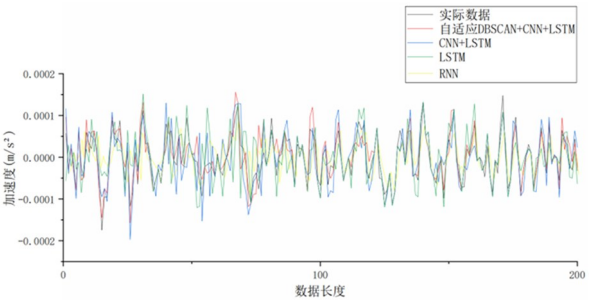


图 17 不同模型预测效果图

Fig. 17 Effect of different model predictions

为更直观地表现出模型预测性能的优劣，通过调研时间序列预测文献选取模型评价指标[17,18]，最终选择模型预测性能的评价指标为平均绝对误(mean absolute error, MAE)和均方根误差(root mean square error, RMSE)，计算公式如式(13)和式(14)所示.

$$E_{\mathrm{MAE}}=\frac{1}{N}\sum_{i=1}^{N}\left|X_{\mathrm{pred}}(i)-X_{\mathrm{act}}(i)\right| \tag{13}$$

$$E_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[X_{\mathrm{act}}(i)-X_{\mathrm{pred}}(i)\right]^{2}} \tag{14}$$

其中，N 表示预测序列的长度，X_{pred} 和 X_{act} 分别表示模型的预测值和实际值，RMSE 反映预测值与实际值的平均偏差，其取值一般大于等于 0，且越接近 0 说明模型预测精度越高；MAE 的取值同样是越接近 0 说明模型预测性能越好[19]. 对比结果如图 18 所示. 可知相比于其余模型，自适应 DBSCAN-CNN-LSTM 组合模型能够有效提高预测精度，不论在 MAE 还是 RMSE 上都低于其余模型，由此表明该方法在桥梁监测系统中的时间序列数据预测方面更有优势.

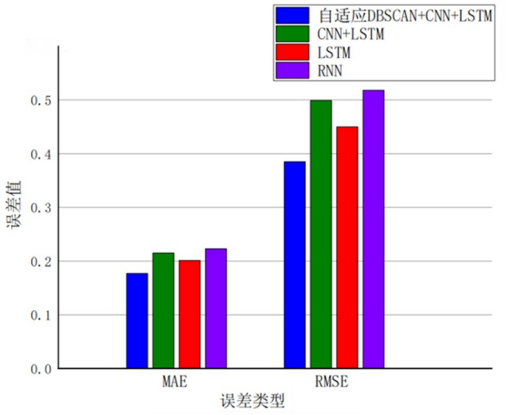


图 18 不同模型预测结果对比

Fig. 18 Comparison of different model prediction results

4.3.4 数据异常检测结果

通过计算得知预测误差，将预测误差的 95% 中位数设为阈值 u，超过其部分的样本用来估计 GPD 参数，并计算 90% 置信水平下的异常阈值. 随着样本的逐渐输入，定期更新训练集以及模型参数，可实现对数据的实时动态监测，如图 19 所示.

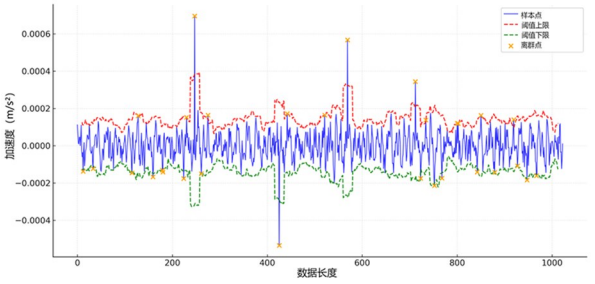


图 19 数据的动态阈值控制图

Fig. 19 Dynamic threshold control charts for data

从表 5 可知，自适应 DBSCAN-CNN-LSTM 组合模型算法在五种数据异常检测中整体检测准确率、精确率、召回率和 F_1 分数均优于 CNN-LSTM、LSTM 和 RNN 方法，是因为该算法将聚类和深度学习相结合，提高了异常检测的准确性和模型的泛化能力，使其不局限于单一异常类型，在发生多种异常下保持对其准确有效的识别能力，但也使其预测耗时略长于单独使用 LSTM 或 RNN 模型. RNN 和 LSTM 模型在整体异常识别效果上

差异不大,是由于 LSTM 在处理大量数据中,易遗忘前面部分的数据特征,导致预测精度随着数据的持续输入而逐渐降低,进而影响阈值的计算,最终导致数据异常检测精度下降.

表 5　整体传感器异常数据检测方法评估

Table 5　Evaluation of overall sensor anomaly data detection methods

模型选择	评估指标				
	准确率(A)	精确率(P)	召回率(R)	F_1 分数($F_{1-\mathrm{Score}}$)	预测耗时/s
自适应 DBSCAN-CNN-LSTM	0.973	0.935	0.921	0.927	258
CNN-LSTM	0.948	0.919	0.885	0.902	279
LSTM	0.894	0.825	0.816	0.820	245
RNN	0.906	0.821	0.797	0.809	239

5　结论

桥梁健康监测异常数据成因复杂,具有体量大、随机性、间歇性等特点.对不同类型的异常数据进行识别,能够保障桥梁结构的安全性与可用性.本文提出一种基于自适应 DBSCAN-CNN-LSTM 组合预测模型的桥梁健康监测数据异常检测方法,并利用公开的 Dowling Hall 加速度监测数据验证该方法的有效性.主要结论如下:

(1)所提出的组合模型能够高效、准确地识别出由传感器故障所造成的数据异常,为桥梁实时预警提供了可靠的保障.

(2)改进的 DBSCAN 密度聚类异常检测方法是以核密度估计理论为基础,无需获知数据分布的先验以及全局信息.其参数的自适应选取也避免了人工选取的局限性.

(3)提出的方法采用无监督学习方式,通过对结构健康状态下的数据建模,避免了训练集类不平衡问题,提高异常数据检测的效率.

参考文献

[1] 单德山,罗凌峰,李乔. 桥梁健康监测 2020 年度研究进展[J]. 土木与环境工程学报(中英文),2021,43(增刊 1):129－134.
SHAN D S, LUO L F, LI Q. State-of-the-art review of the bridge health monitoring in 2020 [J]. Journal of Civil and Environmental Engineering, 2021, 43(S1): 129－134. (in Chinese)

[2] 孙利民,尚志强,夏烨. 大数据背景下的桥梁结构健康监测研究现状与展望[J]. 中国公路学报,2019,32(11):1－20.
SUN L M, SHANG Z Q, XIA Y. Development and prospect of bridge structural health monitoring in the context of big data [J]. China Journal of Highway and Transport, 2019, 32(11): 1－20. (in Chinese)

[3] 《中国公路学报》编辑部. 中国桥梁工程学术研究综述·2021 [J]. 中国公路学报,2021,34(2):1－97.
Editorial Department of China Journal of Highway and Transport. Review on China's bridge engineering research: 2021 [J]. China Journal of Highway and Transport, 2021, 34(2): 1－97. (in Chinese)

[4] 罗浩恩. 桥梁结构健康监测系统传感器自诊断方法研究[D]. 重庆:重庆大学,2016.
LUO H E. Research on fault diagnosisof sensors for bridgestructural health monitoring system [D]. Chongqing: Chongqing University, 2016. (in Chinese)

[5] 胡顺仁,陈伟民,章鹏,等. 基于关联分析的传感器连续失效数据识别研究[C]//2008 中国仪器仪表与测控技术报告大会. 北京:中国仪器仪表学会,2008:80－82.
HU S R, CHEN W M, ZHANG P, et al. Research of continuous invalid data identification based on multi-informaton correlation analysis [C]//2008 China Instrumentation and Measurement & Control Technology Progress Conferenc. Beijing: China Instrument and Control Society, 2008: 80－82. (in Chinese)

[6] 袁慎芳,梁栋,高宁,等. 基于结构健康监测系统的桥梁数据异常诊断研究[J]. 电子科技大学学报,2013,42(1):69－74.
YUAN S F, LIANG D, GAO N, et al. The bridge data diagnosis research based on structural health monitoring system [J]. Journal of University of Electronic Science and Technology of China, 2013, 42(1): 69－74. (in Chinese)

[7] JEONG S, FERGUSON M, HOU R, et al. Sensor

data reconstruction using bidirectional recurrent neural network with application to bridge monitoring [J]. Advanced Engineering Informatics, 2019, 42: 100991.

[8] SMARSLY K, LAW K H. Decentralized fault detection and isolation in wireless structural health monitoring systems using analytical redundancy [J]. Advances in Engineering Software, 2014, 73: 1－10.

[9] FU Y G, PENG C, GOMEZ F, et al. Sensor fault management techniques for wireless smart sensor networks in structural health monitoring [J]. Structural Control and Health Monitoring, 2019, 26(7): e2362.

[10] BAO Y, TANG Z Y, LI H, et al. Computer vision and deep learning-based data anomaly detection method for structural health monitoring [J]. Structural Health Monitoring, 2018. 18(2): 401－421.

[11] ESTER M, KRIEGEL H P, XU X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise [C]// International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM, 1996: 226－231.

[12] 董晓君，程春玲. 基于核密度估计的 K-CFSFDP 聚类算法[J]. 计算机科学，2018，45(11)：244－248.
DONG X J, CHENG C L. K-CFSFDP clustering algorithm based on kernel density estimation [J]. Computer Science, 2018, 45(11): 244－248. (in Chinese)

[13] 王康，周治平. 高斯核密度估计方法检测健康数据异常值[J]. 计算机科学与探索，2019，13(12)：2094－2102.
WANG K, ZHOU Z P. Gaussian kernel density estimation method for detecting abnormal values of health data [J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(12): 2094－2102. (in Chinese)

[14] SANDER J, ESTER M, KRIEGEL H P, et al. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 169－194.

[15] YI T H, HUANG H B, LI H N. Development of sensor validation methodologies for structural health monitoring: a comprehensive review [J]. Measurement, 2017, 109: 200－214.

[16] MOSER P, MOAVENI B. Design and deployment of a continuous monitoring system for the Dowling hall footbridge [J]. Experimental Techniques, 2013, 37(1): 15－26.

[17] GUO A P, JIANG A J, LIN J, et al. Data mining algorithms for bridge health monitoring: Kohonen clustering and LSTM prediction approaches [J]. The Journal of Supercomputing, 2020, 76(2): 932－947.

[18] LIU J W, LI Q, CHEN W R, et al. Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks [J]. International Journal of Hydrogen Energy, 2019, 44(11): 5470－5480.

[19] 靳启文. 基于结构鲁棒性分析能量方法与数据挖掘技术的桥梁结构健康监测应用研究[D]. 合肥：合肥工业大学，2019.
JIN Q W. Application study on bridge structural health monitoring with consideration of structural roubustness analysis energy method and data mining technology [D]. Hefei: Hefei University of Technology, 2019. (in Chinese)