基于深度学习的无线传感网络高维数据异常检测方法

胡挺峰
（无锡城市职业技术学院，江苏 无锡 214153）

摘要：数据异常检测是计算机算法领域的一个重要课题，为提高检测精度，基于深度学习算法设计一种无线传感网络高维数据异常检测方法。获取浅层的自编码器，得到低维向量和高维向量的差别，在编码器中传递隐藏函数，均方误差的权重，避免训练过拟合，信息传播过程中的代价函数，建立多个数据异常节点的差异化矩阵，基于基于深度学习算法实现无线传感网络的数据分类。设计高维数据异常检测算法，得到异常检测结果。实验数据表明，在训练数据比例不同的情况下，数据比例越大，检测精度越高。该方法在五个数据集中的准确率、召回率、F1 值均在 0.98 以上，可见其具备较高的检测精度，且适用性较强。

关键词：深度学习；无线传感网络；高维数据；数据异常检测
中图分类号：G647　　**文献标识码：**A　　**文章编号：**2096-9759(2023)06-0087-03

0 引言

在大数据时代到来以后，无论是在数据的精密处理领域还是数据应用领域，均需要对海量高维数据进行检测。但是，数据的浪潮汹涌而至，每时每刻都有大量的信息产生，数据的维度也每时每刻都在增加，从最初的几维成长到几十维、上百维。很多算法都受困于传统的数据处理难题，集中于针对静态数据的解算，而忽略了对高位数据中关键信息的提取与应用。在现有的相关研究中，文献[1]结合时间序列异常数据的修正与检测技术，对数据进行高精度的提取与分析，通过建立回归模型的方式，解决了梯度消失等问题，并对误差序列的估计进行修正与管理，结合更深层次的网络结构，获取了更高精度的检测结果。文献 [2] 将变压器的运行和维护作为实验对象，结合已有的监测数据，剔除了其中大量的伪数据，并获取了一种基于凝聚层次聚类算法的时间序列检测模型。在该模型的基础上，确定数据类型，并通过实时监测的数据获取异常结果。文献[3]为解决高维传感器中的数据受到环境扰动影响较大的问题，获取了一阶差分信号序列的关键节点，并将空间中具备相关性的传感器实时划分到同一簇类，使用分割算法预设特征属性，获取了异常检测的判定结果。结合上述文献，本文设计了一种基于深度学习的无线传感网络高维数据异常检测方法。

1 基于深度学习算法建立无线传感网络数据分类模型

为获取无线传感网络数据中的分类信息，需要建立一个分类模型。首先需要使用训练及建立隔离森林，随即划分部分节点特征，在重构误差函数之后，将特定的神经网络作为解决表征向量的功能指标，同时利用三层结构建立编码网络，通过深度学习方法获取的浅层自编码器[4-5]。

在该自编码器中，编码结构与解码结构均由连接层获取，

收稿日期：2023-02-03
作者简介：胡挺峰（1973.07-），男（汉族），江苏无锡人，硕士，副教授，主要研究方向：信息安全，数据分析，神经网络等。

原始 U-Net 模型相较，在网络性能、模型识别率等方面都取得显著的提高，提取的道路结构更具有完整性、连通性。下一步研究方向可以放在模型精度优化与道路形态相结合方面。

参考文献：

[1] 戴激光,王杨,等.光学遥感影像道路提取的方法综述[J].遥感学报,2020,24(07):804-823.
[2] 贾建鑫,孙海彬,等.多源遥感数据的道路提取技术研究现状及展望[J].光学精密工程,2021,29(02):430-442.
[3] 左娟, 李勇军. 结合纹理与形状特征的高分辨率遥感影像道路提取[J].测绘,2013,36(03):111-113.
[4] 阙昊懿,黄辉先,等.基于双阈值 SSDA 模板匹配的遥感图像道路边缘检测研究[J].国土资源遥感,2014,26(04):29-33.
[5] 谢志伟,平继伟,等. 基于邻域特征的电子地图道路交叉点自动提取[J].中国科技论文,2020,15(5):599-604.
[6] 朱芳芳,李仲勤,等.特征分量的城市建筑物面向对象提取方法[J].测绘科学,2020,45(01):84-91.
[7] Long J, Shelhamer E, et al. Fully convolutional networks for semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition,2015: 3431-344.
[8] Olaf Ronneberger;Philipp Fischer; et al.U-Net:Convolutional Networks for Biomedical Image Segmentation [J]. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015,2015,Vol.9351: 234-241
[9] Badrinarinarayanan V, Kendall A, et al. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12):2481-249
[10] Chen L C, Papanantidreou G, et al. Rethinking atrous convolution for semc image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.
[11] 林娜,张小青,等.空洞卷积 U-Net 的遥感影像道路提取方法[J].测绘科学,2021,46(09):109-114+156.
[12] 张新华,黄梦醒,等.基于深度学习的卫星图像道路分割算法[J].计算机工程,2021,47(10):306-313.
[13] 陈洪云,孙作雷,孔薇.融合深度神经网络和空洞卷积的语义图像分割研究[J].小型微型计算机系统,2020,41(01):166-170.
[14] 宋廷强,刘童心,等.改进 U-Net 网络的遥感影像道路提取方法研究[J].计算机工程与应用,2021,57(14):209-216.
[15]Q. Hou, D. Zhou and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition.11

且连接层有且仅有一个，这种编码结构通常用于对网络高维数据进行降维处理。在降维的过程中，未标记的数据通常仅仅具备有代表性的特征，而编码器的结构则需要通过函数从其中的一段输入到隐藏层，经过解码器的处理后再回流到原始数据特征中[6-7]。在上述模型中，左边为编码器，右边为与编码器相对应的解码器，中间部位则为隐藏层。不同的向量之间存在低维向量和高位向量的差别，在传递隐藏函数的过程中，可以得到：

$$f(x)=g^{(1)}\left(w^{(1)}x+b_i\right) \tag{1}$$

式中，$f(x)$ 表示编码器由输入层到隐藏层的映射函数值；$g^{(1)}$表示编码器的激活函数；$w^{(1)}$表示权重矩阵；x表示编码器函数的自变量；b_i表示浅层编码器的偏差系数。结合估计向量，可以通过损失函数判断训练结果。在重建值与输入值之间，获取均方误差的权重，为避免训练过拟合，公式为：

$$k(w,p)=\frac{\sum_{i=1}^{n}\left(x_i-x_j\right)^2+\lambda_d}{N_m}\times\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{p=1}^{k}\left(w_{ij}^{(p)}\right)^2}{2} \tag{2}$$

式中，$k(w,p)$ 表示正则惩罚下训练过拟合函数；x^i和x_j分别表示样本的第 i 个表示向量和第 j 个表示向量；λ_d表示拟合系数；N_m表示样本总量；$w_{ij}^{(p)}$表示训练数据中隐藏层的权重值。将该编码器中的最小神经元构建为一个整体的神经网络，需要使用深度学习算法进行编码，并得到信息传播过程中的代价函数：

$$f_{\cos t}=\frac{\sum_{i=1}^{n}f(x,y)}{n_p} \tag{3}$$

式中，f_{cost}表示信息传播的代价函数值；$f(x,y)$表示最小化标签误差；n_p表示梯度下降指标。想要尽最大程度提高数据分类的准确性，还需要对编码器进行降噪处理。在自编码器的训练过程中，添加与编码结构相对应的有损数据，对原始数据进行降噪处理，并将其连接在编码与解码的数据解析功能中。为避免过拟合导致的鲁棒性，需要通过非线性的拓展功能，提取包含异常值的高维数据，进而将其中具备隐藏功能的数据进行梯度扩散[13]。针对前文中提到的自编码器模型，则需要使用随机下降或者反向传播的方式，经由数量引发质量的迭代，并将已经传播号对编码器结构与重构结果相连，最后再继续堆叠。为提高网络学习算法的性能，在编码层与解码层之间，增加一个具备重构特性的标准变分隐藏结构，在保持KL散度不变的前提下，完成对输入数据的重构处理。除利用标准模型中的低维数据进行随即分割以外，还可以通过泛化能力，输入数据中的唯一直接投影，使其更具备潜在的密度估计空间。获取多个数据异常节点的差异化矩阵：

$$H_{diff}=\begin{bmatrix}\cos\theta_{11} & \cos\theta_{12} & \mathrm{L} & \cos\theta_{1n}\\ \cos\theta_{21} & \cos\theta_{22} & \mathrm{L} & \cos\theta_{2n}\\ \mathrm{L} & \mathrm{L} & \mathrm{L} & \mathrm{L}\\ \cos\theta_{m1} & \cos\theta_{m2} & \mathrm{L} & \cos\theta_{mn}\end{bmatrix} \tag{4}$$

式中，H_{diff}表示多个数据异常节点的差异化矩阵；$cos\theta_{mn}$表示 m 维度下第 n 个向量的余弦值。就此可以获取高维数据的分类模型。

2 高维数据异常检测算法设计

通过上述分类方法，能够将无线传感网络中的高维数据异常值提取出来，据此可以设计一种针对高维数据异常检测的算法，如图 1 所示。

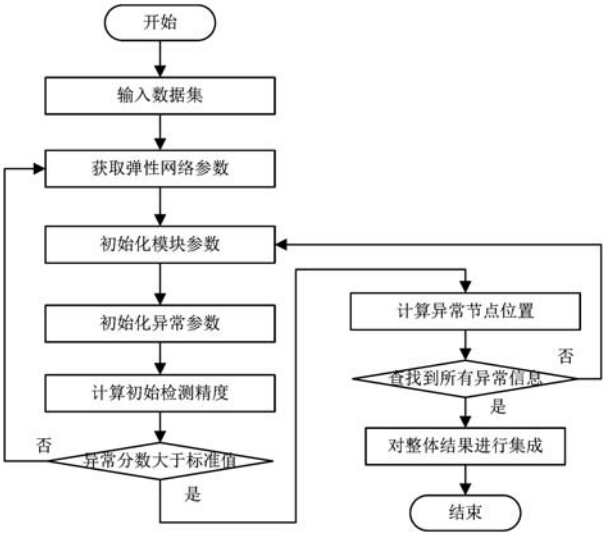

图 1 算法流程

如图 1 所示，首先需要输入数据集，并获取弹性网络参数。对模块参数和异常参数进行初始化处理，执行每一层异常数据的弹性检测程序。在弹性网络中，以线性回归模型作为少数通过正则化处理的回归岭系数，结合稀疏过程，自动选择弹性网络变量，可以得到与之相匹配的变量组信息。数据的原始特征也可以作为某项预测因子，在达到特征提取的预期数值时，利用高维数据的回归分析理论，保证网络参数不会出现过拟合和欠拟合的现象。与传统的数据异常检测方法不同，该检测方法将数据对应的路径长度作为异常分数，在计算路径长度的过程中，就可以直接随即采样，通过与二叉树结构十分相似的平均搜索特性，对样本示例进行定义。计算初始检测精度，判断异常分数是否大于标准值。异常分数在高维数据的异常检测中是一个十分重要的节点，其可以针对原始数据进行候补选择，并在通过弹性模块获取异常特征之后，对异常进行打分，计算路径长度，并对路径进行归一化处理。集成所有异常分数，每一层都会在结束时获取求和结果：

$$N_p=\sum_{i=1}^{n}M_i \tag{5}$$

式中，N_p表示异常分数的求和结果，M_i则表示地 i 个异常分数。在每一层的集成中，如果前一项的均方误差大于上一次检测值，就需要重复该项操作，直至误差检测通过。对异常得分进行单位化处理后，即可得到该层的异常分数。同时计算并判断异常节点的位置，检测该异常信息是否为该数据集中的最后一个异常点。如果不是最后一个异常信息，需要返回到模块参数初始化的环节，重复运行该算法。但是如果异常信息全部查找成功，就需要对整体结果进行集成化处理，得到异常检测结果。

3 实验研究

3.1 实验数据集与实验参数设置

在测试上文中基于深度学习的无线传感网络高维数据异常检测方法的有效性时，需要结合数据集的数据围堵，根据数据对象的数量以及质量，设置实验参数。为保证实验结果的准确性，选择四个数据集作为实验对象。其中的数据集 A 来自于某病例分类样本，数据集 B 来自于某网络入侵数据集，数据集 C 主要来自于垃圾邮件数据以及电力通信数据，数据集 D 主要来自于传感器网络节点。设置实验参数，选择其中的正

常数据作为训练数据，并将数据集中 10 的数据作为测试数据。在不同的数据集中设置不同的分布映射函数，从数据集 A 到数据集 D 分别为 4、8、16、32。

3.2 评估指标

使用平均准确率、平均召回率以及 F1 分数作为该异常检测方法的评估指标，计算公式分别为：

$$P_{re}=\frac{|F_g|\cap|F_r|}{|F_r|} \tag{6}$$

$$R_{call}=\frac{|F_g|\cap|F_r|}{|F_g|} \tag{7}$$

$$F_1=\frac{2P_{re}R_{call}}{P_{re}+R_{call}} \tag{8}$$

式中，P_{re} 表示该高位数据异常检测方法的准确率；P_{call} 表示召回率；F_g 和 F_r 分别表示数据集中的异常集与报告中的异常集；F_1 为一种评估检测精度的指标。根据上述三个公式，可以得到该检测算法的性能。

3.3 实验数据与分析

3.3.1 训练数据比例对实验结果的影响分析

测试不同比例的训练数据对最终算法精度的影响，分别设置数据集中参与训练的数据比例为 20%、40%、60%、80%，得到 4 个数据集的准确率、召回率以及 F1 指标，其结果如下表 1 所示。

表 1 数据参与比例不同对算法性能的影响

指标	比例/%	数据集 A	数据集 B	数据集 C	数据集 D
准确率	20	0.4478	0.4474	0.5141	0.5045
	40	0.5254	0.5321	0.5732	0.6121
	60	0.6156	0.6141	0.6552	0.7351
	80	0.7254	0.7841	0.8515	0.7725
召回率	20	0.3174	0.3145	0.3174	0.3165
	40	0.3915	0.3315	0.4396	0.3441
	60	0.4524	0.3552	0.4924	0.4215
	80	0.5185	0.3714	0.5544	0.4841
F1 值	20	0.3241	0.4141	0.3247	0.3345
	40	0.4163	0.4916	0.5514	0.4162
	60	0.5385	0.5541	0.6543	0.5212
	80	0.6826	0.6585	0.7532	0.6714

如表 1 所示，随着数据比例的增加，相同条件下的算法性能均随之提高。在数据集 A 中，当数据比例训练为 20%时，准确率、召回率以及 F1 值分别为 0.4478、0.3174、0.3241。但是当数据训练比例增加至 80%时，算法的准确率、召回率以及 F1 值已经提高到了 0.7254、0.5185、0.6826。数据集 B、数据集 C、数据集 D 的检测精度判定结果与数据集 A 相同。由此可见，同一个数据集中的数据，随着参与比例的增加，检测精度也会相应提高。

3.3.2 不同数据集下算法精度对比

以上述 4 种数据集为实验对象，分别测试本文深度学习算法以及传统的几种算法的数据异常检测精度，实验结果如表 2 所示。

表 2 不同算法性能对比

评价指标	数据集	深度学习算法	多尺度深度残差网络	层次聚类分析	BIRCH 聚类算法
准确率	A	0.9915	0.9147	0.8985	0.8945
	B	0.9941	0.8585	0.8841	0.8814
	C	0.9852	0.9141	0.8952	0.8952
	D	0.9863	0.8241	0.8832	0.8863
召回率	A	0.9912	0.9452	0.8741	0.8841
	B	0.9941	0.8162	0.8852	0.8942
	C	0.9852	0.8241	0.9115	0.8941
	D	0.9863	0.9152	0.9063	0.8852
F1 值	A	0.9814	0.9412	0.8925	0.9047
	B	0.9952	0.9141	0.8874	0.9085
	C	0.9941	0.8252	0.8785	0.8996
	D	0.9925	0.8135	0.8946	0.8845

对比四种不同的高维数据异常检测方法，在 4 个数据集下，本文算法的准确率、召回率、F1 值均在 0.98 以上。而多尺度深度残差网络方法的三种评价指标则在 0.81-0.95，层次聚类分析算法的评价结果在 0.87-0.92，BIRCH 聚类算法的算法检测精度在 0.88-0.91。由此可见，本文方法的高维数据异常检测精度远高于传统的三种算法。

4 结语

本文设计了一种基于深度学习的无线传感网络高维数据异常检测方法，该方法可以在数据维度逐渐增加的情况下，获取大规模数据的提取与分析结果。实验结果显示，该方法的检测精度远高于其他几种对比算法。在针对高维数据的异常掩盖问题时，则可以将异常数据的构建以及网络特征提取等作为执行条件。在下一步的研究中，可以在网络中输入异常特征向量，并构建多组序列，使用集成学习的算法改变整体的运行效率。

参考文献：

[1] 潘玲玲,刘俊,夏旻.基于多尺度深度残差网络的时间序列异常数据检测与修正[J].计算机应用与软件,2022,39(07):38-43+173.

[2] 王文森,杨晓西,刘阳,等.基于层次聚类分析的变压器油中溶解气体在线监测数据异常检测 [J]. 高压电器, 2023,59(01):142-147.

[3] 赵娇.基于 BIRCH 聚类算法的高维传感器数据异常检测[J].传感技术学报,2022,35(12):1686-1690.

[4] 张靖,孙文举,尼文斌,等.基于深度学习的风洞天平测力试验数据异常检测方法研究[J].实验流体力学,2022,36(06):67-73.

[5] 刘明群,何鑫,覃日升,等.基于改进 K-means 聚类 k 值选择算法的配网电压数据异常检测 [J]. 电力科学与技术学报,2022,37(06):91-99.

[6] 董小瑞,孙伟,樊群才,等.基于 KLDA-INFLO 的继电保护整定数据异常识别方法[J].电力科学与技术学报,2022,37(06):132-137+149.

[7] 程雅琼.基于关联规则的无线通信网络异常数据检测方法[J].长江信息通信,2022,35(04):43-45.