Computer Science and Application 计算机科学与应用, 2024, 14(3), 20-30
Published Online March 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.143054

Hans汉斯

通信基站关键指标数据异常检测与趋势预测的研究

窦钰哲*，甘　俊

西藏大学信息科学技术学院，西藏　拉萨

收稿日期：2024年2月8日；录用日期：2024年3月6日；发布日期：2024年3月18日

摘　要

本文主要研究了智能运维中的异常检测与趋势预测的问题。采用高斯滤波器、归一化处理等方法，完成了对数据的预处理；利用Hampel滤波函数获取异常值，通过快速傅里叶变换计算异常周期，最后使用数据训练模型和决策树分类器，建立基于BP神经网络的滚动预测模型，实现时间序列中的异常预测和趋势预测。

关键词

时间序列数据挖掘，通信基站，BP神经网络，Hampel滤波，决策树分类器

Research on Anomaly Detection and Trend Prediction of Key Index Data of Communication Base Station

Yuzhe Dou*, Jun Gan

College of Information Science and Technology, Tibet University, Lhasa Tibet

Received: Feb. 8th, 2024; accepted: Mar. 6th, 2024; published: Mar. 18th, 2024

Abstract

This paper mainly studies the problems of anomaly detection and trend prediction in intelligent operation and maintenance. Gaussian filter, normalization processing and other methods are used to complete the preprocessing of the data; use the Hampel filter function to obtain abnormal val-

*通讯作者。

文章引用: 窦钰哲, 甘俊. 通信基站关键指标数据异常检测与趋势预测的研究[J]. 计算机科学与应用, 2024, 14(3): 20-30. DOI: 10.12677/csa.2024.143054

ues, calculate the abnormal cycle through fast Fourier transform, and finally use the data training model and decision tree classifier to establish The rolling prediction model based on BP neural network realizes the requirements of abnormal prediction and trend prediction.

Keywords

Time Series Data Mining, Communication Base Station, BP Neural Network, Hampel Filter, Decision Tree Classifier

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

智能运维是将人工智能应用于运维领域，是以人工智能作为主导地位，由 AI 托管运维全程，通过机器学习从而发现和解决传统的自动化运维无法解决的问题。在智能运维中，异常检测、异常预测、趋势预测这三个问题的解决与否，是智能运维[1]能否正常进行的关键。

在 29 天内，对 58 个小区的运营商基站 71 个 KPI 的性能指标每一小时采样一次总共 29 * 58 * 24 = 40,368 条数据。

1) 选取小区内的平均用户数、小区 PDCP 流量、平局激活用户数作为三项核心指标。以这三项指标为关键指标检测出 29 日内全部小区所有的异常数据。

2) 预测未来三天上述三项核心指标的取值。

3) 利用(1)中得到的异常数值，建立预测模型，跟据模型输入时间跨度及输出时间跨度分析模型准确率，预测未来是否会出现异常数值。

2. 基本理论

2.1. Hampel 滤波

Hampel 滤波算法[2]是决策滤波器的一种，用于寻找信号序列中的异常值，并以较为合适的值代替异常值。该滤波器假设给定的数据集服从一个分布和概率模型，然后依据假设采用不一致检验处理信号序列。本质上是一种基于中值和 MAD (median absolute deviation)尺度估计器的离群值检测程序。

具体来说，对输入序列 D, Hampel 滤波器的输出响应如下：

$$Y_k = \begin{cases} D_k\left(\left|D_k - M_k\right| \le TS_k\right) \\ M_k\left(\left|D_k - M_k\right| > TS_k\right) \end{cases} \tag{1}$$

式(1)中的 M_k 是样本及其前 $K-1$ 个样本组成的滑动窗口的中值，定义为

$$M_k = \text{median}\left\{D_{k-(K-1)}, \cdots, D_{k-2}, D_{k-1}, D_k\right\} \tag{2}$$

式(2)中，median 是求给定数值的中值的函数。K 是一个正整数，称为窗口宽度。S_k 是 MAD 尺度估计，定义为

$$S_k = 1.4826 \times \text{median}_{j\in[1,k]}\left\{\left|D_{k-j}, -M_k\right|\right\} \tag{3}$$

式(3)中，“1.4826”是一个工程经验值，它使 MAD 尺度估计成为高斯数据标准偏差的无偏估计。T 是一个动态阈值调优参数设置为 3，即查找与中位数相差超过 3 个标准偏差的样本，则认为该值为异常值。

2.2. BP 神经网络

BP 网络(Back-ProPagation Network) [3]又称反向传播神经网络，通过样本数据的训练，不断修正网络权值和阈值使误差函数沿负梯度方向下降，逼近期望输出。它是一种应用较为广泛的神经网络模型，多用于函数逼近、模型识别分类、数据压缩和时间序列预测等。

典型的 BP 神经网络模型结构如图 1 所示，主要包括输入层、中间层(隐含层)及输出层。其中，中间层可以拓展为多层，相邻层之间通过各神经元实现全连接，而同一层之间的各神经元无连接，其信号从输入层经过中间层流向输出层。其运算流程主要分为两个阶段：样本信号的前向传播和转化，以及误差的逆向反馈。在信号前向传播的过程中，输入信号通过输入层的归一化处理以及隐含层的非线性计算，从输出层产生相应的输出信号。通过与实际输出信号进行对比得到误差，借由误差的反向传播，不断调整网络的权值和阈值，当得到的输出结果在预定的误差范围内时，网络训练结束。

BP 网络由输入层、隐层和输出层组成，隐层可以有一层或多层，下图 1 是 $m \times k \times n$ 的三层 BP 网络模型结构图，网络选用 S 型传递函数：

$$f(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{4}$$

通过反传误差函数：

$$E=\frac{\sum_i \left(T_i+O_i\right)^2}{2} \tag{5}$$

(T_i 为期望输出、O_i 为网络的计算输出)不断调节网络权值和阈值使误差函数 E 达到极小。

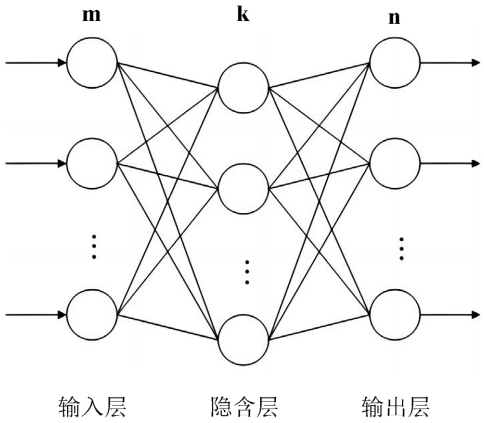

Figure 1. BP neural network model structure
图 1. BP 神经网络模型结构

2.3. 决策树分类器

ID3 作为一种经典的决策树算法[4]，是基于信息熵来选择最佳的测试属性，其选择了当前样本集中具有最大信息增益值的属性作为测试属性。ID3 算法根据信息论理论，采用划分后样本集的不确定性作为衡量划分样本子集的好坏程度，用“信息增益值”度量不确定性——信息增益值越大，不确定性就更小[5]，这就促使我们找到一个好的非叶子节点来进行划分。

假设一个这样的数据样本集 S，其中数据样本集 S 包含了 s 个数据样本，假设类别属性具有 m 个不同的值(判断指标)：$C_i\left(i=1,2,3,\cdots,m\right)$，$S_i$ 是 C_i 中的样本数，对于一个样本集总的信息摘为：

$$I\left(s_1,s_2,\cdots,s_m\right)=-\sum_{i=1}^{m}P_i\log_2 P_i \tag{6}$$

其中，P_i 表示任意样本属于 C_i 的概率，也可以用 s_i/s 进行估计。我们假设一个属性 A 具有 k 个不同的值 $\{a_1,a_2,\cdots,a_k\}$，利用属性 A 将数据样本 S 划分为 k 个子集 $\{S_1,S_2,\cdots,S_k\}$，其中 S_j 包含了集合 S 中属性 A 取 a_j 值的样本。若是选择了属性 A 为测试属性，则这些子集就是从集合 S 的节点生长出来的新的叶子节点。

$$E\left(A\right)=\sum_{j=1}^{k}\left[\frac{s_{1j},s_{2j},\cdots,s_{mj}}{s}\times I\left(s_{1j},s_{2j},\cdots,s_{mj}\right)\right] \tag{7}$$

最后，我们利用属性 A 划分样本集 S 后得到的信息摘增益[6]为：

$$\mathrm{Gain}\left(A\right)=I\left(s_1,s_2,\cdots,s_m\right)-E\left(A\right) \tag{8}$$

3. 模型建立与问题的求解

3.1. 异常检测

3.1.1. 高斯滤波平滑

高斯滤波[7]就是对整幅图像进行加权平均的过程，每一个像素点的值，都由其本身和邻域内的其他像素值经过加权平均后得到。高斯滤波器平滑处理后降低噪声的影响。采用高斯滤波器，系统函数是平滑的，避免了振铃现象。

3.1.2. 归一化处理

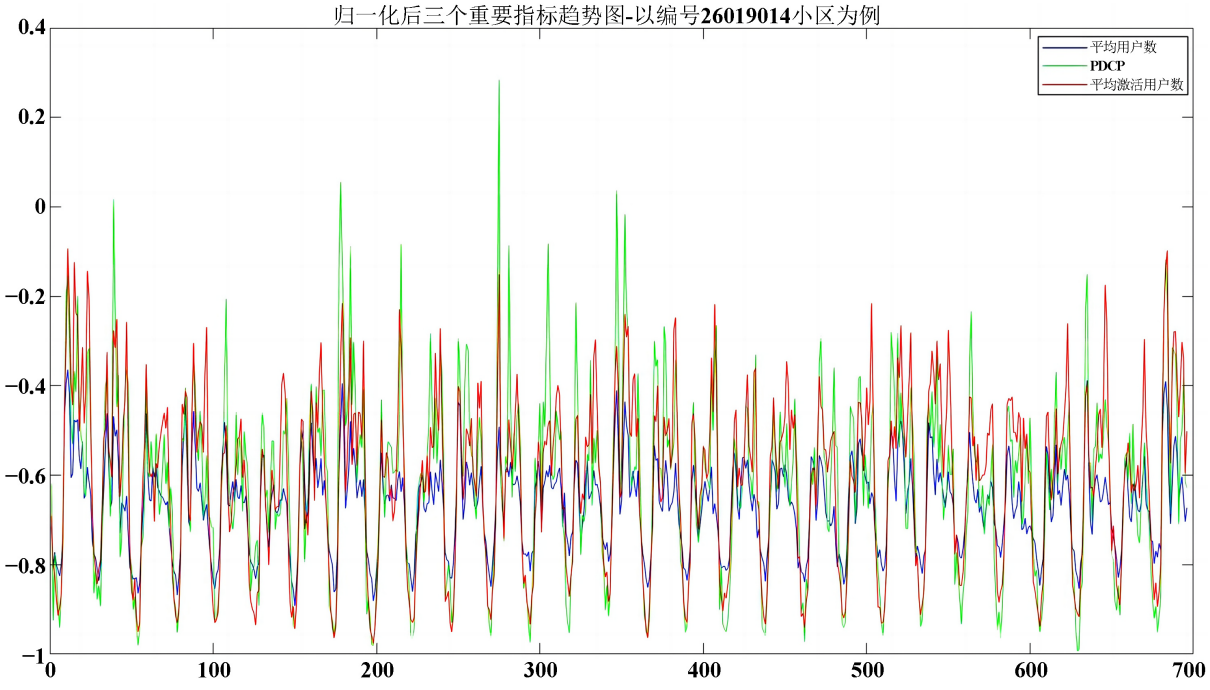

Figure 2. Trend chart of three important indicators after normalization - Take 26019014 as an example
图 2. 归一化后三个重要指标趋势图-以编号 26019014 小区为例

由于该数据是否为异常数据受到多个指标的影响，将各项数据通过归一化处理，使各项指标处于同一量级，消除其它相关变量带来影响，而着重于看三个关键指标带来的影响，便于接下来的综合对比评价[6]。归一化后结果如上图 2 所示。

3.1.3. 皮尔逊相关系数

利用 Matlab 计算输出皮尔逊相关系数[8]，三项指标各自相关性如下：

- 小区内的平均用户数与小区 PDCP 流量的相关性：r = 0.9323
- 小区内的平均用户数与平均激活用户数的相关性：r = 0.9198
- 平均激活用户数与小区 PDCP 流量的相关性：r = 0.9248

据皮尔逊相关系数原理得，对于异常检测中三项指标两两相对，具有高度相关关系。

3.1.4. 使用 Hampel 滤波算法求解

在 Hampel 滤波算法的原理基础上，将经过高斯滤波平滑、归一化处理的指标数据导入 Matlab，绘制原始信号、滤波信号和异常值，标注离群点位置。本文通过将附件 1 的数据导入 Matlab，通过 Hampel 滤波函数进行处理，可以得到如下图 3 所示结果。

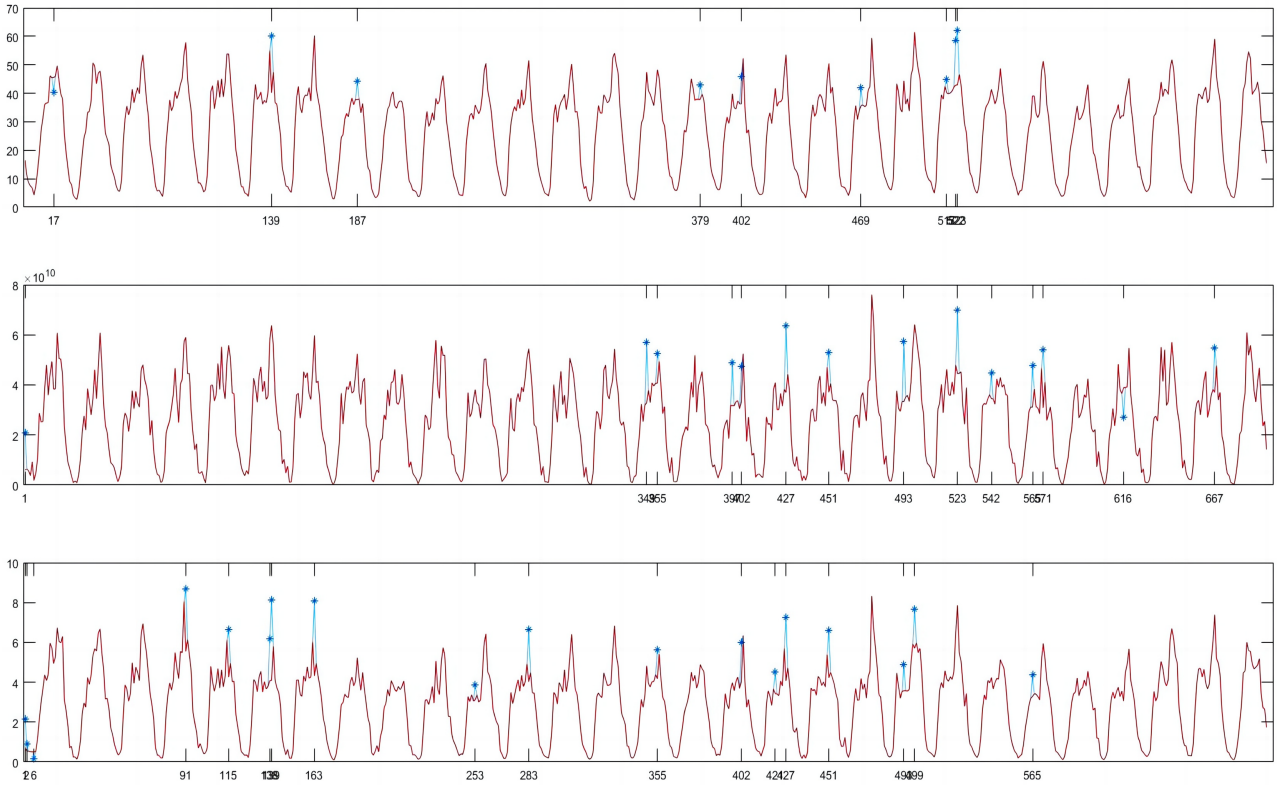

Figure 3. Hampel exception filtering
图 3. Hampel 异常滤波处理

3.1.5. 时间周期的选择

针对各个小区的三项核心指标，根据快速傅里叶变换算法(FFT)和互信息法[9]，计算出各个小区的时间周期；若计算出的时间周期大于 12，就舍弃；否则就保留数值；最后将所有小于 12 的时间周期取平均值，获得时间周期的值。快速傅里叶变换周期功率下图 4 所示。快速傅里叶变换功率下图 5 所示。

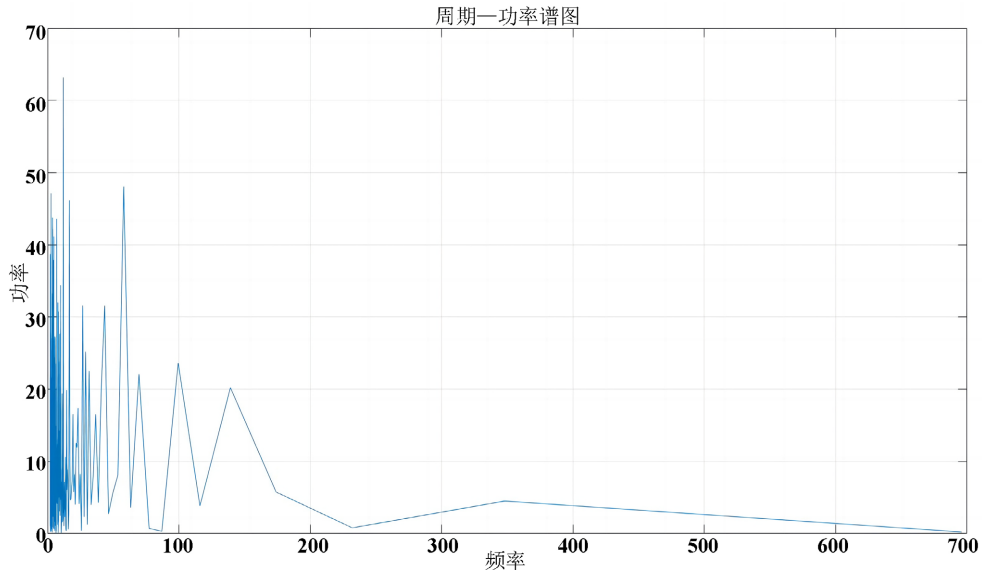

Figure 4. Fast Fourier transform periodic power diagram
图 4. 快速傅里叶变换周期功率图

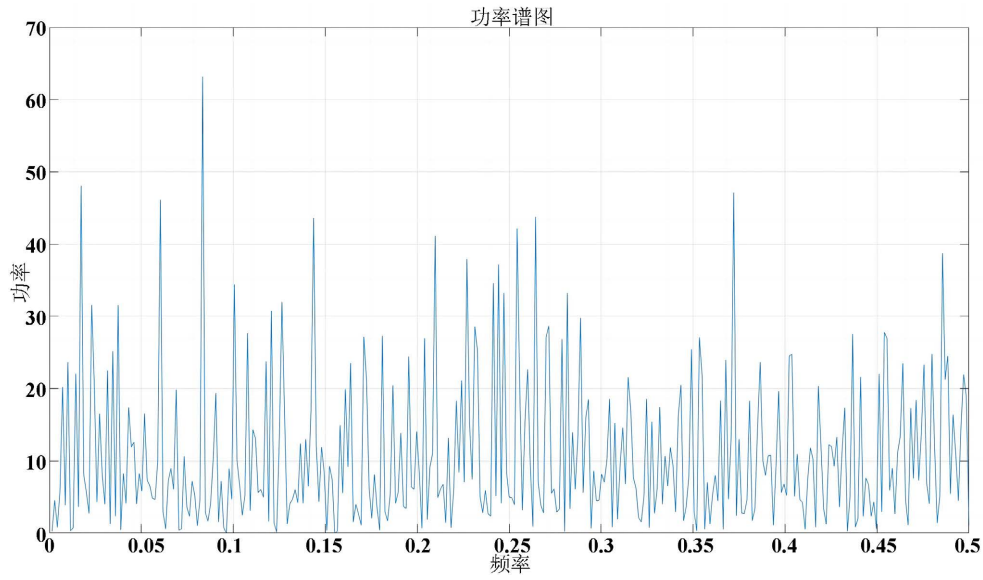

Figure 5. Fast Fourier transform power
图 5. 快速傅里叶变换功率

3.2. 预测未来三天数据趋势

BP 神经网络滚动预测模型[10]的建立采用 Matlab 内置方法 newff 进行建模和训练，经过多次实验，本方案将 newff 的各项参数设定为如下：第一个参数为经过 Hampel 滤波处理后的数据；第二个参数设定隐层神经元的个数为 10 个，输出层的神经元个数为 1 个；第三个参数设定隐层神经元传输函数为“tansig”，输出层传输函数为“purelin”；其余采用默认参数。通过 Matlab 内置的神经网络工具箱用梯度下降法对模型进行训练，实时滚动预测即采用神经网络进行单步预测，用 t 时刻的数据预测 t + 1 时刻数据，t + 1 时刻将采集的实测数据用于 t + 2 时刻的预测，以此类推，滚动的补充新数据，剔除旧数据。实时滚动预测[11]也称为实时跟踪预测，由于每次预测都利用了最新的实测数据，预测的结果也更加准确。预测未来

三天数据趋势下图 6 所示。

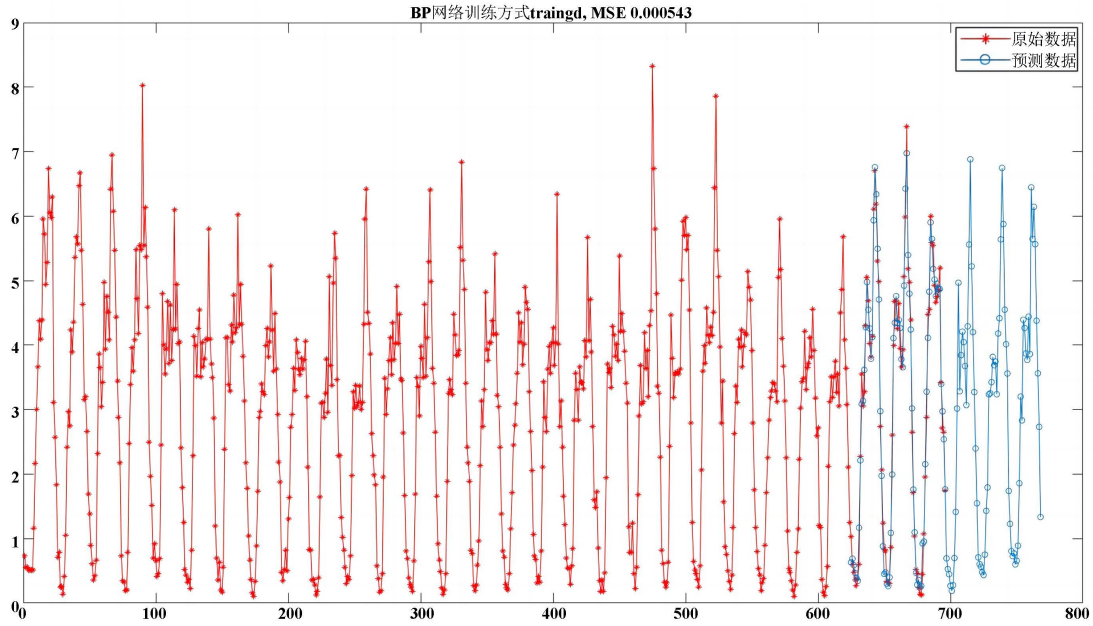

Figure 6. Forecast data trends for the next three days
图 6. 预测未来三天数据趋势

3.3. 异常预测

3.3.1. 自相关系数分析

将三项关键指标数据与异常数据进行自相关性分析[12]。得出 autocorr 结果如下图 7、xcorr 结果如下图 8 所示。

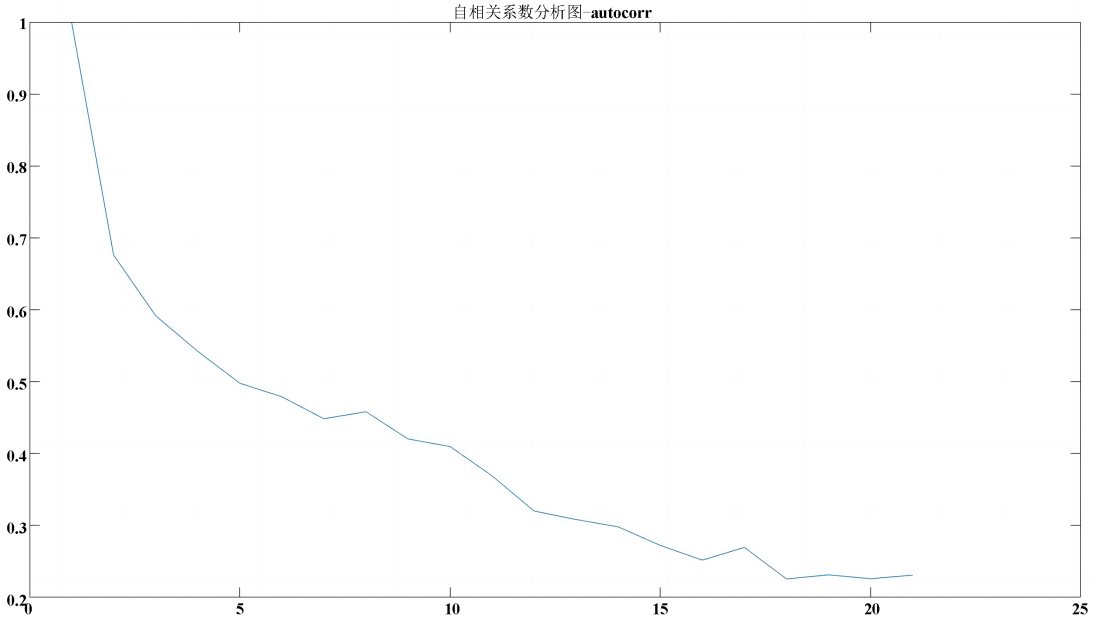

Figure 7. Autocorrelation coefficient analysis chart—autocorr
图 7. 自相关系数分析图——autocorr

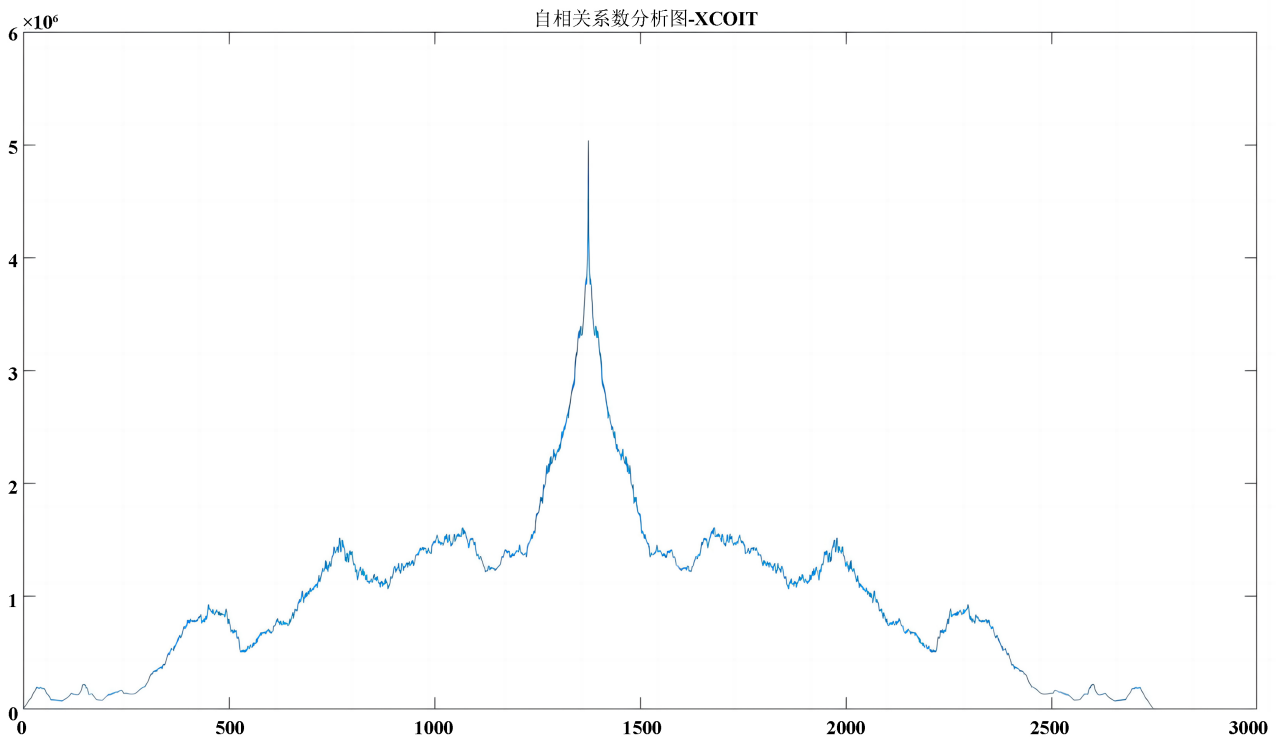

Figure 8. Autocorrelation coefficient analysis chart—xcorr
图 8. 自相关系数分析图——xcorr

3.3.2. 皮尔逊相关系数分析选取指标数据

将三项关键指标数据与异常数据进行皮尔逊相关系数分析，对于相关系数：$R \geq 0.9$，数据选为特征指标，得到与异常数据相关性较高的特征指标。结果分别见下表 1~3。

Table 1. Correlation coefficients of other indicators corresponding to abnormal data of the average number of users in the community
表 1. 小区平均用户数异常数据对应的其他指标相关系数

特性指标下标	相关系数	特性指标下标	相关系数	特性指标下标	相关系数
5	0.9579	6	0.9277	7	0.9277
8	0.9199	9	0.9200	11	0.9214
20	0.9106	22	0.9618	24	0.9471
38	0.9270	39	0.9539	40	0.9551
55	0.9949				

Table 2. Correlation coefficients of other indicators corresponding to abnormal PDCP data in the community
表 2. 小区 PDCP 异常数据对应的其他指标相关系数

特性指标下标	相关系数	特性指标下标	相关系数	特性指标下标	相关系数
20	0.9316	22	0.9520	24	0.9152
39	0.9001	55	0.9339		

Table 3. Correlation coefficient of other indicators corresponding to abnormal data of average activated users
表 3. 平均激活用户数异常数据对应的其他指标相关系数

特性指标下标	相关系数	特性指标下标	相关系数	特性指标下标	相关系数
5	0.9287	8	0.9016	9	0.9016
11	0.9055	20	0.9127	24	0.9033
39	0.9067	40	0.9202	55	0.9595

3.3.3. BP 神经网络滚动预测模型

BP 神经网络滚动预测模型[13]的建立采用 Matlab 内置方法 newff 进行建模和训练，经过多次实验，本方案将 newff 的各项参数设定为如下：第一个参数为经过 Hampel 滤波处理后的数据；第二个参数设定隐层神经元的个数为 10 个，输出层的神经元个数为 1 个；第三个参数设定隐层神经元传输函数为“tansig”，输出层传输函数为“purelin”；其余采用默认参数。通过 Matlab 内置的神经网络工具箱用梯度下降法对模型进行训练，实时滚动预测即采用神经网络进行单步预测，用 t 时刻的数据预测 t + 1 时刻数据，t + 1 时刻将采集的实测数据用于 t + 2 时刻的预测，以此类推，滚动的补充新数据，剔除旧数据。实时滚动预测也称为实时跟踪预测，由于每次预测都利用了最新的实测数据，预测的结果也更加准确。

3.3.4. 决策树分类器异常预测

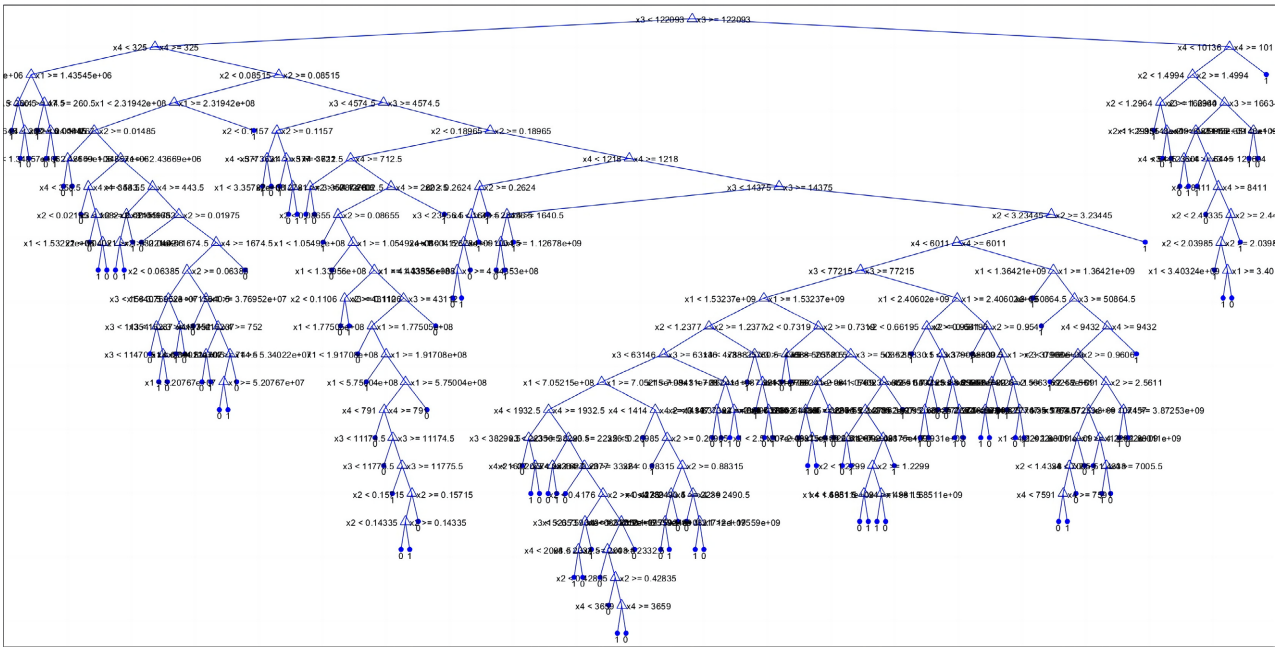

Figure 9. Decision classification tree diagram
图 9. 决策分类树图示

决策树分类技术[14]是数据挖掘中的高效技术，该技术利用属性比较方法获取样本分类结果。因为决策树是采用自顶向下的递归方式，从树的根节点，经过若干个叶子节点，每个叶子节点代表目标类别属性的值，故而我们把经过问题一处理后的数据进行分类，选取部分被问题一判定出的某一指标下的异常数据所占据的一行的所有数据，作为一个叶子节点的输入数据，满足该叶子节点的数据即为异常数据，否则进入下一叶子节点的判定，除了异常数据，本方案还会抽取部分正常数据作为其他叶子节点的数据

[11]，最终异常数据和正常数据的数目比达到 1:1。以上操作借助 Matlab 自带的统计工具箱函数 ClassificationTree.fit (X, Y)，即可完成决策树分类器的建立。在决策树分类器建好后本方案采用 Matlab 自带的统计工具箱函数 eval (t, X)，从正常数据与异常数据并存的混合数据中选取 75%的数据，输入决策树分类器中，对决策树分类器进行训练。最终训练出的决策分类树结构见上图 9。

4. 结果

数据分析

本文通过 BP 神经网络滚动预测模型分别对决策分类器指标小区内用户平均数、决策分类器指标小区 PDCP 流量和决策分类指标平均激活用户数进行检测[15]。检测结果如表 4~6。

Table 4. Test the average number of users in a decision classifier index cell
表 4. 对决策分类器指标小区内用户平均数的检验

数据种类		正常值	异常值	数值确定		错误	正确率
数据总数	2674	1300	1374	正常值确定	274	52	p1 = 84.0491%
训练集数据总数	2006	974	1031	异常值确定	288	54	p2 = 84.2105%
测试集数据总数	668	326	342				

Table 5. The test of PDCP traffic in decision classifier index cell
表 5. 对决策分类器指标小区 PDCP 流量的检验

数据种类		正常值	异常值	数值确定		错误	正确率
数据总数	3454	1700	1754	正常值确定	363	82	p1 = 81.573%
训练集数据总数	2591	1255	1335	异常值确定	351	67	p2 = 83.9713%
测试集数据总数	863	445	418				

Table 6. Test the average number of active users of the decision classifier index
表 6. 对决策分类器指标平均激活用户数的检验

数据种类		正常值	异常值	数值确定		错误	正确率
数据总数	2057	1000	1057	正常值确定	201	38	p1 = 84.1004%
训练集数据总数	1543	761	781	异常值确定	226	49	p2 = 82.1818%
测试集数据总数	514	239	275				

5. 结语

本文数据进行预处理时采用一维高斯函数、归一化处理，一维高斯系统函数是平滑的，避免了振铃现象，归一化处理使各指标处于同一数量级，适合进行综合对比评价。分别对于不同的指标小区内用户平均数小区内用户平均数、小区 PDCP 流量小区 PDCP 流量数、平均激活用户数平均激活用户数正常值确定率为 84.0491%，p1 = 81.573%，p1 = 84.1004% p1 = 84.1004%，异常值确定率为 p2 = 84.2105%，p2 = 83.9713%，p2 = 82.1818%。

建立模型时，选择建立的 BP 神经网络滚动预测模型，该模型属于 BP 神经网络模型，而 BP 神经网络具有较强的非线性拟合能力、容错能力。

基金项目

西藏大学 2022 年大学生创新训练项目(2022XCX083)。

参考文献

[1] 魏吉平, 李静怡, 王凯旋, 等. 一种面向卫星共视授时应用的星站钟差融合方法[J]. 时间频率学报, 2023, 46(2): 85-93. https://doi.org/10.13875/j.issn.1674-0637.2023-02-0085-09

[2] 李麒. 基于 Hampel 滤波和支持向量回归机的土石坝渗流压力预测研究[C]//中国水利学会, 黄河水利委员会. 中国水利学会 2020 学术年会论文集第四分册. 北京: 中国水利水电出版社, 2020: 88-93. https://doi.org/10.26914/c.cnkihy.2020.069481

[3] 王恒, 唐孝国, 郭俊亮. 带通滤波电路故障的 BP 网络诊断方法与仿真设计[J]. 智能计算机与应用, 2022, 12(7): 185-190+201.

[4] 蔡忠林. 基于决策树的电力系统实时动态安全评估方法研究[J]. 能源与环保, 2021, 43(5): 202-207. https://doi.org/10.19389/j.cnki.1003-0506.2021.05.035

[5] 阮晓宏, 黄小猛, 袁鼎荣, 等. 基于异构代价敏感决策树的分类器算法[J]. 计算机科学, 2013, 40(S2): 140-142+146.

[6] Ali, Y.J., Grace, M.A., Williams, O.S., *et al*. (2022) High-Density Surface EMG Signal Quality Enhancement via Optimized Filtering Technique for Amputees' Motion Intent Characterization towards Intuitive Prostheses Control. *Biomedical Signal Processing and Control*, **74**, Article ID: 103497. https://doi.org/10.1016/j.bspc.2022.103497

[7] 蒋贤海, 谢存禧, 邹焱飚. 一种强噪声下的监护信息降噪方法[J]. 华南理工大学学报(自然科学版), 2011, 39(4): 66-69.

[8] Cansu, B., Elisa, M., Per, C., *et al*. (2020) Operational Guidelines for Emissions Control using Cross-correlation Analysis of Waste-to-Energy Process Data. *Energy*, **220**, Article ID: 119733. https://doi.org/10.1016/j.energy.2020.119733

[9] 石振乔. 基于 PPG 信号的血氧检测算法研究[D]: [硕士学位论文]. 海口: 海南大学, 2022. https://doi.org/10.27073/d.cnki.ghadu.2022.000372

[10] 刘高宏, 吴恩启, 闵锐, 等. 基于优化 BP 网络的类矩形盾构偏心刀盘故障预测[J]. 软件导刊, 2020, 19(10): 111-115.

[11] 杨丽萍, 郭宏升. 决策树分类算法在课程成绩预测中的应用[J]. 电子测试, 2022, 36(17): 56-58. https://doi.org/10.16520/j.cnki.1000-8519.2022.17.022

[12] 陈云烁, 符繁强. 关于智能运维中 KPI 异常检测与预测的研究[J]. 信息系统工程, 2023(9): 118-121.

[13] Liu, X.B., Pan, Y.H., Yan, Y., *et al*. (2022) Adaptive BP Network Prediction Method for Ground Surface Roughness with High-Dimensional Parameters. *Mathematics*, **10**, Article 2788. https://doi.org/10.3390/math10152788

[14] 宋丽. 基于决策树的组合分类器的研究[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2012.

[15] 陈津, 谢辉, 唐胜飞, 等. 基于广义回归模型的充电桩运行异常预测方法[J/OL]. 电测与仪表: 1-8. http://kns.cnki.net/kcms/detail/23.1202.th.20220121.1613.003.html, 2023-10-06.