

基于特征提取和机器学习的异常数据识别算法

赵荣欣^{1,2}, 贾鹏飞^{1,2}

(1.上海市建筑科学研究院有限公司,上海市 201108; 2.上海市工程结构安全重点实验室,上海市 200092)

摘要: 结构健康监测系统的大力发展每天都在产生大量的监测数据。对于结构健康监测系统来说,判断这些产生的监测数据是否正常是对结构健康状态进行分析的第一步,也是关键的一步。同时,监测数据的异常与否也是判断传感器、采集设备、传输设备是否正常工作关键性依据。对于一段数据进行识别,判断数据是属于什么样的异常,是一个多分类的问题。采用基于特征提取和机器学习相结合的算法,对时序数据进行分类,能够快速判断数据是否异常和异常的类型。

关键词: 结构健康监测;机器学习;异常识别;K近邻算法;多分类

中图分类号: TU196

文献标志码: A

文章编号: 1009-7716(2023)08-0250-03

0 引言

近些年来,随着城市的快速发展和建设水平的提高,建设了大量的结构健康监测系统。这些检测系统每天都会产生大量的监测数据。怎么对这些监测数据进行分析处理、获取结构本身的健康状态是结构健康监测领域的难题^[1-2]。

监测的异常数据通常按照识别数据的长度分为个别点的局部异常数据和时序数据的连续异常数据。局部异常数据通常指个别数据发生明显偏离、缺失、超出设备量程等明显不符合真实情况的数据。连续异常数据指的是一段时间的数据有明显的漂移、缺失、极大极小值、方波、多个离群值等。

监测数据的异常判断可以分为单分类和多分类问题。单分类问题是对该数据是否异常做出分类。多分类既要回答是否异常的问题,也要回答是什么种类的异常的问题。相比较,多分类的问题更难,也更符合工程的需要。

对于单一的某种类型的数据异常,可以很方便地通过方差、极值等统计特征或者拟合的某些参数进行判断。比如,对于数据的偏移、极大极小值,可以根据直方图特征判断;对于数据缺失,可以根据采样频率判断;对于数据漂移,可以根据线性拟合的线性项参数进行判断;等等。但是由于数据的复杂性和不可预测性,很难具体罗列出所有的病害特征,并且

有些病害的特征很相似,也很难给出具体的判断指标。姜旭宝等^[3]提出了一种基于变宽直方图的全局异常数据检测方法,通过数据聚类的方式将网络中的动态感知数据聚类成变宽的直方图来准确检测出异常数据。袁慎芳等^[4]使用了17维特征向量采用One-class SVM进行了传感器跳变信号、传感器装机噪声信号、船撞和台风引起的结构异常信号4种异常信号,以及正常加速度信号的识别。范时泉等^[5]使用了15维特征向量采用SVM方法进行风速和加速度两种特征的正常/异常判别。

机器学习算法是一类从数据中自动分析获得规律,并利用规律对未知数据进行预测的算法。很早就有把机器学习算法应用数据的异常判断的研究和应用。王雷等^[6]使用支持向量机法的方法对单个测点的异常数据进行了二分类的识别。胡石等^[7]使用神经网络方法通过固定长度的已测数据预估单个温度数据,将数据分为异常和正常两类进行预测。潘轶彪等^[8]利用神经网络的多层感知器模型,结合滚动学习预报机制,提出一类利用历史数据建模,估计当前数据,并根据估计值与实际测量值间的差异来检测异常值的方法。

1 算法介绍

1.1 数据异常分类

根据时域数据的异常的不同特征,可以将加速度数据分为表1所示7类。

1.2 特征提取

不同的数据异常类型有着不同的数据特征,可以通过数据的特征来分辨不同的数据异常类型。对于某

收稿日期: 2022-07-27

基金项目: 上海市科委优秀技术带头人项目(20XD1432400)

作者简介: 赵荣欣(1971—),男,博士,教授级高级工程师,从事桥梁与隧道工程检测评估工作。

表 1 加速度数据异常分类

序号	异常	描述
1	正常数据	正常的加速度曲线
2	缺失	大部分数据或者全部数据缺失
3	极小值	幅值在小范围内波动,数据无效
4	异常点	多个明显不符合规律的异常点
5	方波	幅值在一个极大值附近和一个极小值附近来回波动
6	趋势	包含明显不符合加速度数据规律的趋势
7	漂移	幅值是非平稳的,包含随机的漂移

种单一的数据异常,可以采用某一个特征将数据进行区分;对于多分类问题,则需要同时提取多个特征进行分析。

常见的数据特征包括:

(1)极值、分位值,包括原始数据和标准化数据的极大值、极小值、中值、各分位值特征,可以获取数据的分布特征。

(2)均值、标准差,可以获取数据的均值、标准差的直观统计特征。

(3)拟合的参数,包括拟合的参数、残差的范数,可以获取数据的趋势项特征。

(4)直方图统计数量,按照直方图将数据划分为若干个数据段,统计每个数据段的数据量占总数据量的比例,可以获取数据的分布特征。

(5)频率特征,包含识别得到的频率和对应的幅值,将数据经过傅里叶变换计算得到对应的频率和幅值,可以获取数据的频域信息。

1.3 机器学习

机器学习算法是通过从数据中获取规律并对未知数据进行预测的算法。常用的机器学习算法有 K 近邻算法、逻辑回归、决策树、支持向量机等。

K 近邻算法通过距离最近的 K 个点的类别来进行分类。距离可以选择欧氏距离、曼哈顿距离等,可以通过调整 K 的大小和给不同距离的点赋予不同的权重来优化算法。K 近邻算法是一种简单高效的机器学习算法。

逻辑回归是一种回归模型,经常用于解决二分类问题,计算量小,速度很快,存储资源低。它可以通过一对一或者一对多的方法来解决多分类的问题。

决策树模型是一种以形式表达的预测分析模型,决策树由结点和有向边组成。结点由内部结点和叶节点组成。内部节点表示一个特征或属性,叶节点表示一个类。

支持向量机的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面,可以解决高维问题和小样本机器学习问题,泛化能力比较强。

1.4 总体思路

数据的分类流程如图 1 所示,将原始信号提取特征,获取一个 n 维的特征向量,然后采用机器学习算法进行分类模型的训练,获取分类模型。



图 1 数据分类流程

2 数值分析

2.1 数据描述

选择国内一个斜拉桥的桥梁结构健康监测系统的 38 个加速度传感器。这些加速度传感器的采样频率为 20 Hz,每个小时的数据为 72 000 个点,作为一个原始信号。选取一个月 31 d 的数据,这样一共有 28 272 个原始信号。

将所有信号进行按照 7 种分类进行标注,如图 2 所示,共得到 13 575 个正常信号、2 942 个缺失信号、1 775 个极小值信号、527 个异常点信号、2 996 个方波信号、5 778 个趋势信号和 679 个漂移信号。样本的分布虽然不是很平均,正常信号样本的数量远多于其他样本,异常点信号和漂移信号样本量较少,但是最少的样本也有 527 个,满足一般机器学习算法的样本数量,可以进行分类和提取。

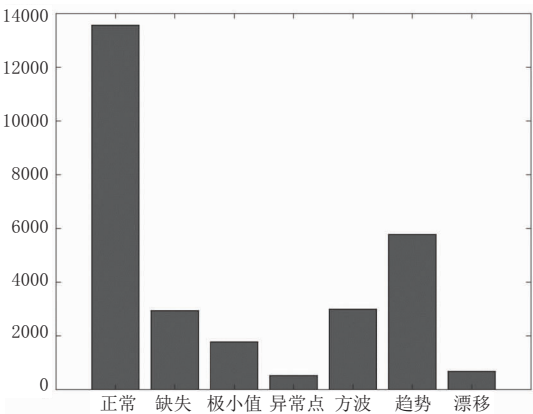


图 2 各类数据的样本数量

选取其中部分加速度信号绘图,如图 3 为一段正常的加速度信号,图 4 为一段最小值异常的加速度信号。

2.2 特征提取

选择原始数据和标准化数据的极大值、极小值、25 分位值、中位值、75 分位值、均值、标准差、线性拟

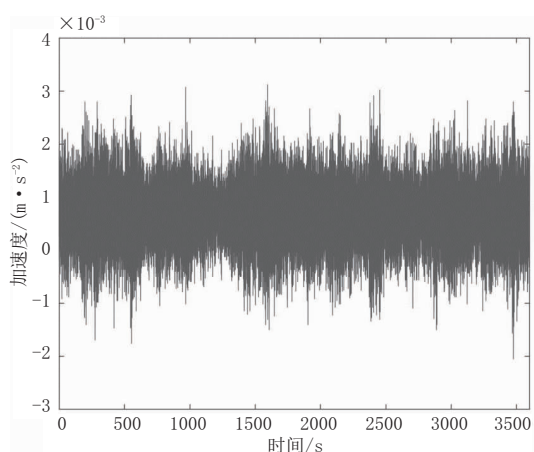


图3 一段正常的加速度信号

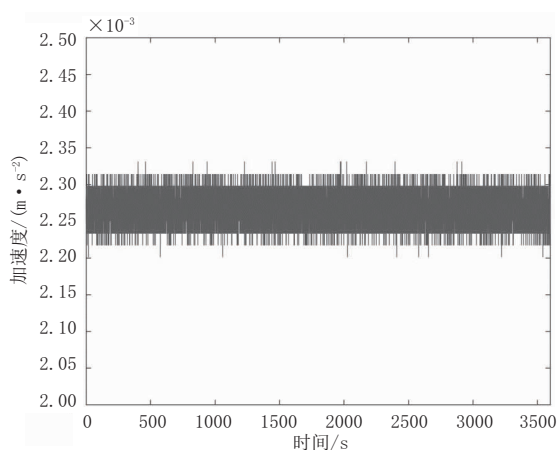


图4 一段极小值异常的加速度信号

合的截距、斜率、残差范数、直方图统计分布特征、频率和频率对应幅值共 21 个特征作为特征向量,将 72 000 维的时域数据转换为 21 维的特征向量。

第 16 个特征为统计分布特征,将第 16 个特征的样本的分布绘图如图 5 所示,单个特征的分布与数据异常的类型并无明显的关系。

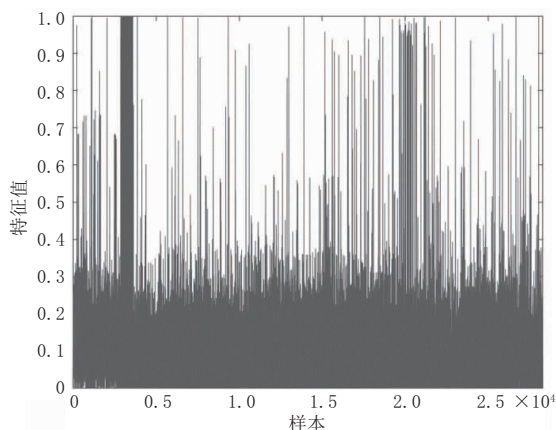


图5 所有样本特征 16 的分布图

3 机器学习

将数据分为 80% 的训练集和 20% 的测试集,采

用 K 近邻算法、逻辑回归多分类算法、决策树算法和支持向量机算法在训练集上进行有监督学习的训练,然后在测试集上进行测试,训练结果见表 2。

表2 训练结果

算法	准确率 /%
K 近邻(网格搜索)	0.97
逻辑回归	0.90
支持向量机	0.83
决策树(最大深度 12)	0.99

训练的结果显示,K 近邻算法和决策树算法在基于特征提取的识别准确率很高,分别达到了 97% 和 99%,而逻辑回归和支持向量机算法的识别结构稍差。

4 结 语

本文提出了一种基于特征提取和机器学习的算法用来解决时域信号的多分类问题,并将其在某桥梁的加速度信号进行了算法的试验和验证。

在真实信号中,K 近邻算法和决策树算法的准确率高达 97% 以上,其他两类算法也能达到 83% 和 90% 的准确率,能够很好地对信号进行分类。

采用这种数据多分类方法,可以快速地对健康监测系统采集的信号进行分类,提取正常信号进行监测指标的计算和预警;而对于异常信号,可能根据异常的种类进行进一步的分析、处理或者预警,对健康监测系统的数据预处理具有一定的意义。

参考文献:

- [1] 孙利民,孙智,淡丹辉,等.我国大跨度桥梁结构健康监测系统研究与应用现状[C]//第十七届全田桥梁学术会议论文集.北京:人民交通出版社,2006: 663. 670.
- [2] 姜绍飞.结构健康监测导论[M].中国:科学出版社,2013.
- [3] 姜旭宝,李光耀,连朔.基于变宽直方图的无线传感器网络异常数据检测算法[J].计算机应用,2011(3): 694-697.
- [4] YUAN S F, LIANG D, GAO N, et al. The bridge data diagnosis research based on structural health monitoring system [J]. Dianzi Keji Daxue Xuebao/Journal of the University of Electronic Science and Technology of China, 2013, 42(1): 69-74.
- [5] 范时泉,张金辉,张其林.结构健康监测系统的异常数据识别[J].计算机辅助工程,2016(5): 60-65.
- [6] 王雷,张瑞青,盛伟,等.基于支持向量机的回归预测和异常数据检测[J].中国电机工程学报,2009, 29(8): 92-96.
- [7] 胡石,李光辉,卢文伟,等.基于神经网络的无线传感器网络异常数据检测方法[J].计算机科学,2014(z2): 208-211.
- [8] 潘轶彪,袁景淇,朱凯,等.基于多层感知器的异常数据实时检测方法[J].上海交通大学学报,2011, 45(8): 1226-1229.