螺丝拧紧曲线异常检测研究说明书

、

修改记录

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 修改人 | 修改日期 | 版本 | 修改内容 |
| 1 | 孟繁库 | 2019-04-11 | V1.0 | 文档创建 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[1. 研究概述 1](#_Toc6227407)

[1.1 文档概述 1](#_Toc6227408)

[1.2 研究概述 1](#_Toc6227409)

[2. 研究过程 1](#_Toc6227410)

[2.1 数据处理 1](#_Toc6227411)

[2.1.1 数据标准化方法 1](#_Toc6227412)

[2.2 聚类算法 3](#_Toc6227413)

[2.3 基于神经网络的曲线分类 5](#_Toc6227414)

[3. 研究展望 5](#_Toc6227415)

# 研究概述

## 文档概述

本文件作为“螺丝拧紧曲线异常检测研究”设计说明，编写的目的是充分叙述本研究取得的进展，以便使用者了解本研究设计思想，实现方法，以及使用范围方法，并为后续的研究工作提供必要的信息。该系统利用机器学习及深度学习技术对螺丝拧紧曲线进行自动化异常检测工作。

## 研究概述

时序异常检测通常形式化为根据某种标准或正常信号寻找离群数据点。该研究利用信息手段对螺丝拧紧曲线数据进行自动化检测，判断异常与否。首先使用采样的方式对样本进行特征归一化，然后使用机器学习中的聚类技术自动划分样本，形成不同的类别，并通过聚类质量评价指标选取合适的聚类数。聚类后结合业务对数据打标签。最后使用深度学习方法，实现螺丝拧紧曲线为输入，曲线所属类别为输出的异常检测功能。研究的整体流程如图1：



图1 螺丝拧紧曲线异常分析研究流程

获取数据后，对数据进行归一化操作，比如剔除无效数据，长度归一化等。然后通过多种相似性度量方法加权的方式构建相似性矩阵，作为谱聚类的输入。对于聚类的结果，需要通过手工的方式剔除部分离群样本，保证有监督学习样本的有效性。最后构建LSTM网络并训练分类模型。

# 研究过程

## 数据处理

### 数据标准化方法

1. 数据截取：如图2所示，螺栓从开始旋入螺母，到其下端面接触被联接连接件，这个过程螺栓不受力，驱动力矩为零，因此采集的数据会有在开始的很长一段时间内力矩的值较小，曲线平稳。在拧紧结束时，力矩突然下降到接近0。对于这两个特殊阶段，通过选取阈值进行截取，以获取数据的有效部分。

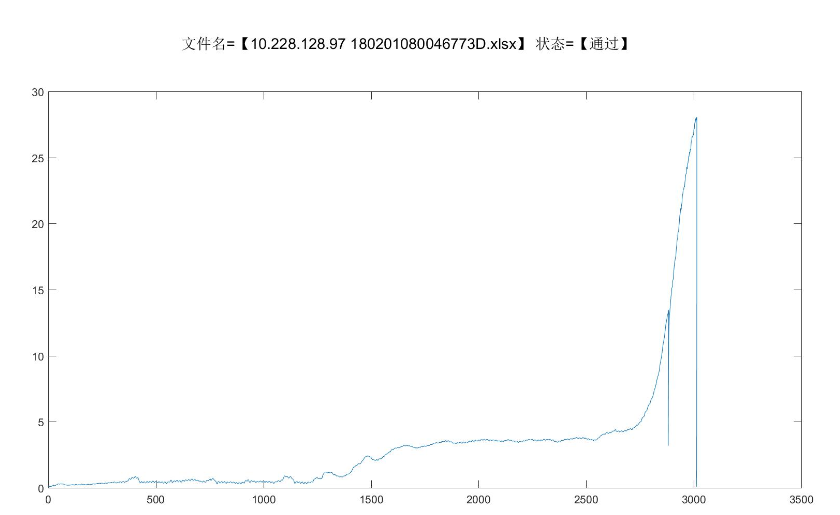


图2 螺丝拧紧曲线示意图

1. 长度归一化：由于采集的数据特征维度不一致，因此需要做长度归一化操作。主要通过两个方式实现，分别是采样以及自动编码机。

采样：首先根据数据长度以及采样后长度确定采样间隔，对于每个间隔，使用均值作为采样后的点，虽然会有信息损失，但是可以保留原始数据的主要特征，截取数据以及采样后的数据如图3所示。

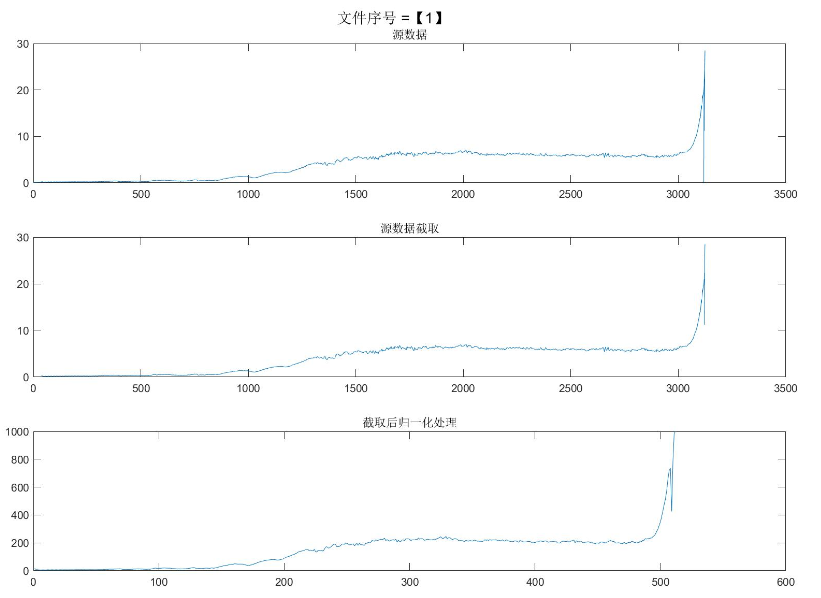


图3 原图经过采样，长度变为512

自动编码机：自动编码机是一个神经网络, 它的意义在于让网络尽可能复现输入信号。为了实现这种复现，自动编码器就必须捕捉可以代表输入数据的最重要的因素。因此可以通过训练自动编码机，获取数据在网络中的隐藏层表示，作为原始数据的编码结果。但是自动编码机的输入需要固定长度，因此对于长度不够的数据，需要补0操作，这会改变原始数据的特征分布，通过实验验证，模型学习到的隐藏层数据表示的并没有忽略补0部分的数据，因此对于需要补0的数据，不适合用自动编码机做长度归一化。

1. 最大-最小规范化：最小-最大规范化也称为离散标准化，是对原始数据的线性变换，将数据值映射到[0, 1]之间。
2. z-score规范化：零-均值规范化也称标准差标准化，经过处理的数据的均值为0，标准差为1。

## 聚类算法

数据特征提取完之后，把数据集做聚类(无监督学习)打标签，并观察输出聚类结果，调整参数直到聚类结果达到较好的效果。本研究尝试了KMeans，SOM以及谱聚类三个方法，经过实验比较，谱聚类的聚类效果要优于另外两个方法。下面重点介绍谱聚类在螺丝拧紧曲线上的聚类应用。

1. 谱聚类介绍

谱聚类（spectral clustering）是广泛使用的聚类算法，比起传统的K-Means算法，谱聚类对数据分布的适应性更强，聚类效果也很优秀，同时聚类的计算量也小很多，实现起来也不复杂。

谱聚类是从图论中演化出来的算法，后来在聚类中得到了广泛的应用。它的主要思想是把所有的数据看做空间中的点，这些点之间可以用边连接起来。距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。

算法流程

输入：样本集D=(x1,x2,...,xn)，相似矩阵的生成方式, 降维后的维度k1, 聚类方法，聚类后的维度k2

输出： 簇划分C(c1,c2,...ck2).

1）根据输入的相似矩阵的生成方式构建样本的相似矩阵S

2）根据相似矩阵S构建邻接矩阵W，构建度矩阵D

3）计算出拉普拉斯矩阵L

4）构建标准化后的拉普拉斯矩阵D−1/2LD−1/2

5）计算D−1/2LD−1/2最小的k个特征值所各自对应的特征向量f

6) 将各自对应的特征向量f组成的矩阵按行标准化，最终组成n×k1维的

特征矩阵F

7）对F中的每一行作为一个k1维的样本，共n个样本，用输入的聚类方法进行聚类，聚类维数为k2。

8）得到簇划分C(c1,c2,...ck2).

1. 相似性矩阵构建

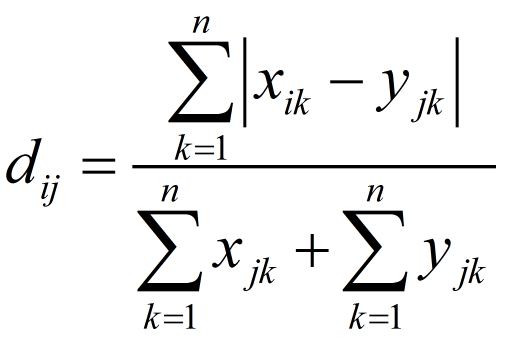
谱聚类的聚类效果依赖于相似性矩阵，因此相似性度量方法的选择尤其重要。

我们采用了三种相似性度量方法如下：

皮尔森系数：用来反映两个变量线性相关程度的统计量。定义为两个变量之间的协方差和标准差的商。

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D283/sign=f0f606054fa7d933bba8e37b9e4bd194/b21c8701a18b87d6d83a970e0a0828381f30fda3.jpg

布雷克蒂斯距离：计算坐标之间的距离。该距离取值在[0,1]之间。



小波分解结合模态经验分解：小波分解使用了pywt中的wavedec模块，对于每个样本做两层分解，获取第一层数据后再进行模态经验分解。模态经验分解使用了pyemd中的emd\_samples模块。

在三种方法单独处理后，以不同的权重对处理结果加权求和，作为最终的相似性度量结果。目前的权重分配采用了均等分配，即每一个方法的权重相同。权重分配要结合曲线的特点以及实验效果调整，不同的相似性度量方法侧重比较不同的特征，对于适合衡量曲线差异性的某一个方法，可以尝试调大其权重。

构建好的相似性矩阵作为谱聚类的输入。Scikit-learn(sklearn)是机器学习中常用的第三方模块，对常用的机器学习方法进行了封装，包括回归(Regression)、降维(Dimensionality Reduction)、分类(Classfication)、聚类(Clustering)等方法。我们使用其中的SpectralClustering模块完成聚类。

聚类加速：构建相似性矩阵的过程的时间复杂度是O(n2)，当样本数量比较大时，程序运行时间较长，因此需要做性能优化。目前采用的方法是多进程加速，同时计算相似性矩阵的每一行，最终将计算后的结果拼接成完整的矩阵。优化性能首受限于机器的物理性能。

聚类评价：对于聚类结果的检测，目前通过可视化操作完成，还不够客观而且需要人力参与，因此需要自动化的检测指标评价聚类质量。可以从簇内的稠密程度和簇间的离散程度来评估聚类的效果。常见的方法有轮廓系数Silhouette Coefficient和Calinski-Harabasz Index等。

## 基于神经网络的曲线分类

通过聚类过程，我们获取了带有标签的样本。有监督学习对于样本的标签的精确度要求较高，因此聚类后的结果要通过手工筛选，剔除部分不符合本类特征的样本。手工剔除后，可以进行有监督学习，训练可以自动分类的模型。

训练平台使用了Keras。Keras是一个高层神经网络API，Keras由纯Python编写而成并基Tensorflow、Theano以及CNTK后端。Keras 为支持快速实验而生，具有简易和快速的原型设计（keras具有高度模块化，极简，和可扩充特性），支持CNN和RNN，或二者的结合以及无缝CPU和GPU切换等特点。

网络架构：采用多层LSTM神经网络。LSTM是一种特殊的RNN模型，常用于处理时序序列，变长序列。由于螺丝拧紧曲线也是时序数据，因此可以采用LSTM提取曲线的特征。网络结构采用了多层LSTM连接的形式，输入层是LSTM层，输出层是Dense全连接层，输出层的神经元个数是分类数。损失函数采用的是mae，优化器为adam。在训练过程中，对于二分类我们发现Dense层的输出值为接近1和接近0的两个值，这样不利于比较曲线所属不同类别的概率。由于是分类问题，因此最后一层可以改为softmax层，损失函数可以更改为交叉熵损失，效果会更好。

在小规模样本上尝试了二分类的训练，训练结果的准确率能达到90%以上。可以增大训练集或者增加分类数，再做训练以比较神经网络在不同规模数据的分类效果。

# 研究展望

目前，螺丝拧紧曲线的基本研究流程已经确定，主要分为数据归一化，聚类，特征提取以及分类几个流程。基于该流程，进行了一些实验验证，实验表明目前的方法具有一定的可行性，但是还需要更多的实验数据衡量算法的可行性，比如增大数据集，尝试对不同的工位数据进行分类，以及网络结构的改进包括分类器以及优化器的选择等。