无标题

- 项目过程
  - 准备阶段
    - 阅读文献
      - 异常检测领域综述性论文（Kerr）
        - DEEP LEARNING FOR ANOMALY DETECTION: A SURVEY
          - **关键字解读**
            - 异常检测（也称为异常值检测）
              - 目标是以数据驱动的方式确定所有与众不同的异常数据。
              - 异常数据和奇异值的区别
            - 深度学习
              - 机器学习的子集
                - 优点
                  - 性能好
                  - 灵活性较高
                  - 数据规模增加时，深度学习的效率高于传统的机器学习
          - **深度异常检测的动机与挑战**
            - 传统方法的性能不够好，因为它无法捕获数据中的复杂结构
            - 传统方法的可扩展性较差，而数据集却是越来越大
            - 传统方法的适应性在减弱，因为正常行为和异常行为的边界越来越模糊，而且这在不断发展
            - 传统方法对手动特征工程依赖性较高，DAD具有自动特征学习功能
          - **分类**
            - **按输入数据**
              - 序列
                - 音频序列
                - 蛋白质序列
                - 时间序列
              - 非序列
                - 图像
            - **按是否监督**
              - 监督式

        - 加标签
      - 半监督式
        - 仅单标签
      - 无监督
        - 无标签
    - **按异常的分类**
      - 点异常
      - 集体异常
        - 单个出现没有问题，成群结队出现就可能是异常
      - 上下文异常
    - 按输出分类
      - 异常分值
        - 给出一个异常程度，然后域专家根据经验设置阈值来判定是否是异常
      - 标签
  - **应用**
    - 入侵检测
      - 网络
      - 主机
    - 欺诈检测
      - 银行
      - 网络
    - 恶意软件检测
    - 医疗异常检测
    - 社交异常检测
      - 垃圾邮件
      - 网络骗子
      - 虚假用户
      - 谣言散布
      - 性侵犯
    - 日志异常检测
    - 物联网数据异常检测
    - **时间序列异常检测**
  - **现有模型**
    - 时空网络模型（STN）
      - 一般的深度学习只能分析时间或空间的特征

        - CNN分析空间特征
        - LSTM分析时间特征
    - 总和产品网络（SPN）
    - 词向量模型（Word2vec）
    - 生成模型
        - 变异自编器（VAE）
        - 生成对抗网络（GAN）
    - 卷积神经网络（CNN）
    - 序列模型
        - RNN循环神经网络
- **针对时间序列异常检测的研究性论文**
    - Profile
        - **Matrix Profile I: All Pairs Similarity Joins for Time Series:A Unifying View tha Includes Motifs, Discords and Shapelets（晓楠）**
            - Matrix Profile
                - 由UCR（加州大学河滨分校）提出的一个时间序列的分析算法。通过一个时间序列，可计算出它的MP，表示的是子序列之间的潜在关系（各种距离）MP也是一个向量（时间序列）。
                - **Shapelets**
                    - 代表一类可以在分类场景中提供直接的解释性和见解的时间序列子序列[6]，并且基于Shapelet的模型在各种研究中都被证明是有前景的
                    - 太难了，相关资料并不多，学姐研究了几周后放弃了这个方向转向和我一起研究TSI
    - Time Series to Image
        - **TSI: Time series to imaging based model for detecting anomalous energy consumption in smart buildings（Kerr）**
    - Others
        - **A Novel Technique for Long-Term Anomaly Detection in the Cloud（2014 引153） (yuan)**
            - 引言
                - 云计算随着发展在现代社会中占据着越来越重要的地位，文章的作者是twitter的一名工程师，他提到twitter当下面临的一个问题是如何自动检测云平台上的长期异常，即**long-term anomalies**。
                - 为此作者在ESD(generalized Extreme Studentized Deviate)的基础上建立了一种新的统计学方法，称之为分段中值法(**picewise median**)，并在实际生数据上测试了其性能。
                - ESD：检测单变量异常值的一种统计学方法，数学表达式如下：

- **公式基础**
  - 时间序列X是x_t的集合，其中t=0,1,2...
  - R_X=X-S_X-T_X
  - 解释：X可分解为三部分
    - Sx是周期性成分，描述序列的周期性变化
    - Tx是趋势成分，描述序列的总体趋势（非周期性）
    - Rx是X去掉Sx和Tx后的剩余成分
  - **异常检测主要针对Rx进行检测**
- **新技术：分段中值法**
  - 周期性成分比较容易确定，趋势成分(trend)如何提取将直接影响结果的好坏，所以如何确定trend十分关键。
  - 经过实际实验发现，以中值代替得到的trend从X中减去，这种方法的性能好于直接减去Tx，能减少误报的产生，如下图：
  - 但是这种方法只对trend比较平的序列比较有效，在有显著trend的长期序列表现很差。
  - 作者考察了现有的两种提取trend的方法：STL trend和分位数回归法(quant regression)，然后提出了新技术——分段中值法，并比较了它们的性能。
  - **STL trend 概要**
    - 从序列中减去估计的trend，将得到的新序列划分为子周期序列(sub-cycle series)
    - 对每个子周期序列运用LOESS(局部加权回归)，估计Sx，从原始序列中去Sx得到新的序列，对新序列运用LOESS得到新的trend估计值
    - 重复上述步骤直到新的trend估计值不再改变或改变得很小
    - 注：局部加权回归LOESS
    - 局部加权回归：以一个点x为中心，向前后截取一段长度为frac的数据，对于该段数据用权值函数w做一个加权的线性回归。对于所有的n个数据点则可以做出n条加权回归线，每条回归线的中心值的连线则为这段数据的Lowess曲线。
  - **分位数回归法**
    - 分位数回归是把分位数的概念融入到普通的回归里，所谓的0.9分位数回归，就是希望回归曲线之下能够包含90%的数据点，这也是分位数的概念。
    - 用B-Spline曲线做分位数回归已被证实是很有效的提取trend的方法，但是这种方法适用于两周以上的长序列，当时间小于两周时会过拟合，这意味着假如出现大块的异常数据隔断了正常数据，这种方法的性能会受严重影响。
  - **分段中值法**

- **上述三种方法的比较**
  - 三种方法提取trend的示意图如下
  - 有效性比较
  - 实时性能比较
  - 注入分析
    - 实际中很难采集到异常数据，可以采用异常注入的方法进行测试
- **A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data（Brian）**
  - 2019年 AAAI会议发表
  - 本论文主要解决了多变量时间序列数据中的异常检测中的几个问题：
    - 时间依赖性问题，能捕获在不同时间步长中的时间依赖关系
    - 实际应用中的噪声对模型预测结果影响较大，此模型尽可能将噪声的影响降到最低
    - 在现实应用中，可以根据异常事件的严重程度，在异常检测的基础上进行**异常评分**，分析出现异常的根本原因
  - 本论文的主要贡献
    - 提出了一种多尺度卷积循环编解码器（MSCRED），来进行多变量时间序列数据的异常检测和诊断，较好地解决了上述三个问题
    - 签名矩阵
      - 描述不同时间步长中系统状态的多个级别，可以表示不同变量的时间序之间的相互关系，级别就是用来表示异常事件的严重程度
    -
      - 使用**卷积编码器**对时间序列编码
      - 并开发了基于注意力的ConvLSTM网络来捕获时间模式ConvLSTM被开用于捕获视频序列中的时间信息，但其性能可能会随着序列长度的增加恶化。
      - 加入了注意力机制（Attention Based ConvLSTM）后，可以跨不同的时间步自适应地选择相关的隐藏状态。
      - 利用**卷积译码器**重构特征矩阵，并用残差特征矩阵进行异常检测和诊断
  - 实验结果
    - 数据集：合成数据集、电厂数据集
    - 与其他模型比较，MSCRED模型的异常检测性能是最佳的
    - 根本原因检测
      - 主要对比了MSCRED和LSTM-ED的性能，MSCRED将所有时间序列按其异常得分进行排序，将得分最高的k个序列作为根本原因。
      - 结果显示，MSCRED的表现比LSTMED高出约25.9%。

- **Outlier Detection for Time Series with Recurrent Auto-encoder Ensembles（young young）**
  - 论文主题与贡献
    - 解决某些情况下自动编码器对异常值过度拟合造成对整体质量的影响
  - 论文中提出的理论模型
    - 集成自编码器
      - 目的与优势
        - 减小由于单个自动编码器产生的偶然误差所带来的误差
        - 提高基于自编码器的异常检测的准确性
      - 实际使用
        - 对每个自动编码器随机删除一些连接得到稀疏自编码器
        - 降低整体重建误差的方差，提高结果准确性
      - 缺点与劣势
        - 只能用于非序列数据
        - 不能直接用于基于时间序列的异常检测
    - S-RNN集成自编码器
      - IF独立框架
        - 模型描述
          - 集成包含多个S-RNN自编码器
          - 每个自编码器由一个编码器和一个解码器组成
          - 每个自动编码器有其不同的稀疏权值向量
        - 模型结构
    - SF共享框架
      - 共享目的
        - 充分考虑各自编码器加的关联于共性
      - 模型描述
        - 每个自编码器模块分别重构原始时间序列
        - 使用共享层来完成各模块间的交互
      - 模型结构
      - 共享优势
        - 共享层中的参数起到使共享状态稀疏的作用
        - 避免过度拟合，提高系统鲁棒性
  - 论文中的相关实验
    - 使用数据集
      - 单变量数据集

                - NAB
            - 多变量数据集
                - ECG
        - 对照实验方法
            - 现有解决方案
                - LOF、SVM
                - ISF、MP、RN
                - CNN、LSTM
        - 实验的具体实现
            - Python 3
            - Tenserflow 1
            - Scikit-learn 1
        - 评价机制
            - 评价原则
                - 不依赖具体阈值
                - 反映真阳性、真阴性、假阳性、假阴性全面权衡
            - 具体机制
                - PR-AUC
                - ROC-AUC
        - 实验结果
            - 本文提出模型与既有模型的比较
                - 深度学习方法效果更优
                - 本文方法较其他集成自编码器更适用于序列数据
                - 集成方法优于大多数单独方法
            - 自编码器数量的影响
                - 自编码器数量增加，结果更优
            - 本文两种集成方法的比较
                - 共享框架性能更好
                - 独立框架内存消耗较少
- 整理数据集（Kerr）
    - REFIT 电气负载测量数据集
    - 电力负荷图2011-2014数据集
    - Numenta异常基准 (NAB) 数据集
    - 加利福尼亚交通运输数据集 (PeMS数据集)
    - Yahoo's Webscope S5 数据集
    - 2018 AIOps's KPI-Anomaly-Detection 数据集

- ECG数据集

搜索模型

- 邱博模型（yuan）
  - 数据集
    - aiopsdata
    - 单变量
  - 模型概述（**ConvLSTM**）
    - 构建
    - ConvLSTM就是在LSTM之前加卷积操作，邱博的模型架构为三层卷积池化+LSTM+softmax
    - 训练时，训练数据以窗口的形式传到模型里进行训练
  - 运行结果
- CNN模型（yuan）
  - CNN简述
    - 框架结构
      - 卷积层【提取特征】
      - 池化层【降维，减少运算，避免过拟合】
      - 全连接层【分类】
    - 常见运用
    - **CNN用于时间序列异常检测**
  - 模型介绍
- RNN模型（yuan）
  - **结构**：RNN中的每个节点都有关联，如下图所示，Xt表示t时刻的输入，Ot是t时刻对应输出， St是t时刻的存储记忆。对于RNN中的每个单元，输入分为两个部分：
    - 1）当前时刻的真正的输入Xt；
    - 2）前一时刻的存储记忆St-1。
  - **常见运用**：RNN 常用于序列是相互依赖的（有限或无限）数据流，所以适合时间序列的数据，它的输出可以是一个序列值或者一序列的值。
  - **在时间序列检测中的应用**
    - 数据集
      - **ECG变量**
      - **gesture(双变量)**
      - **nyc_taxi(三变量)**

- **power_demand(单变量)**
- **respiration(单变量)**
- **space_shuttle(单变量)**

- 模型构建及训练

- 针对ECG数据的TSI模型（Kerr）
  - Github https://github.com/giorgiodema/ECG-Anomaly-Detection
  - Dataset
  - Data process
  - Model
  - Result
  - Requirements

- 实施阶段
  - 构建模型（Kerr）
    - 在服务器上配置环境
    - 训练模型
    - 测试模型
  - 设计前端（young young）
    - 前端设计框架
    - 设计功能点
  - 前后端连接（young young）
    - 使用框架
    - 数据交互技术
- 测试阶段(Undo)
  - 功能测试
  - 性能测试
- 发布阶段(Undo)
  - 整理代码
  - 完善文档

- 实用工具
  - Microsoft Todo
  - 幕布
  - Postman
  - Github