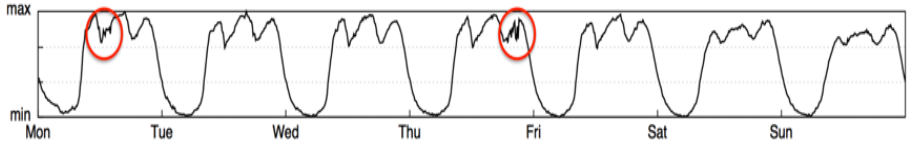
基于机器学习的时间序列自动化异常检测

队伍：D.I. 领队：李富 （队员：葛帅，曹海峰）2018.05.19

KPI(时间序列)异常检测

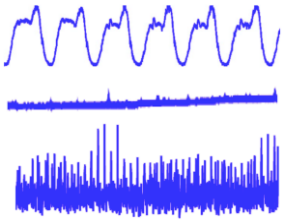
挑战问题：



- **时间序列**：根据时间顺序测量的数值组成的数列
- **KPI异常检测**：通过**算法**分析时间序列数据，判断其是否出现异常行为。本质是数据的**二分类**问题。

问题特点：

KPI三种形态



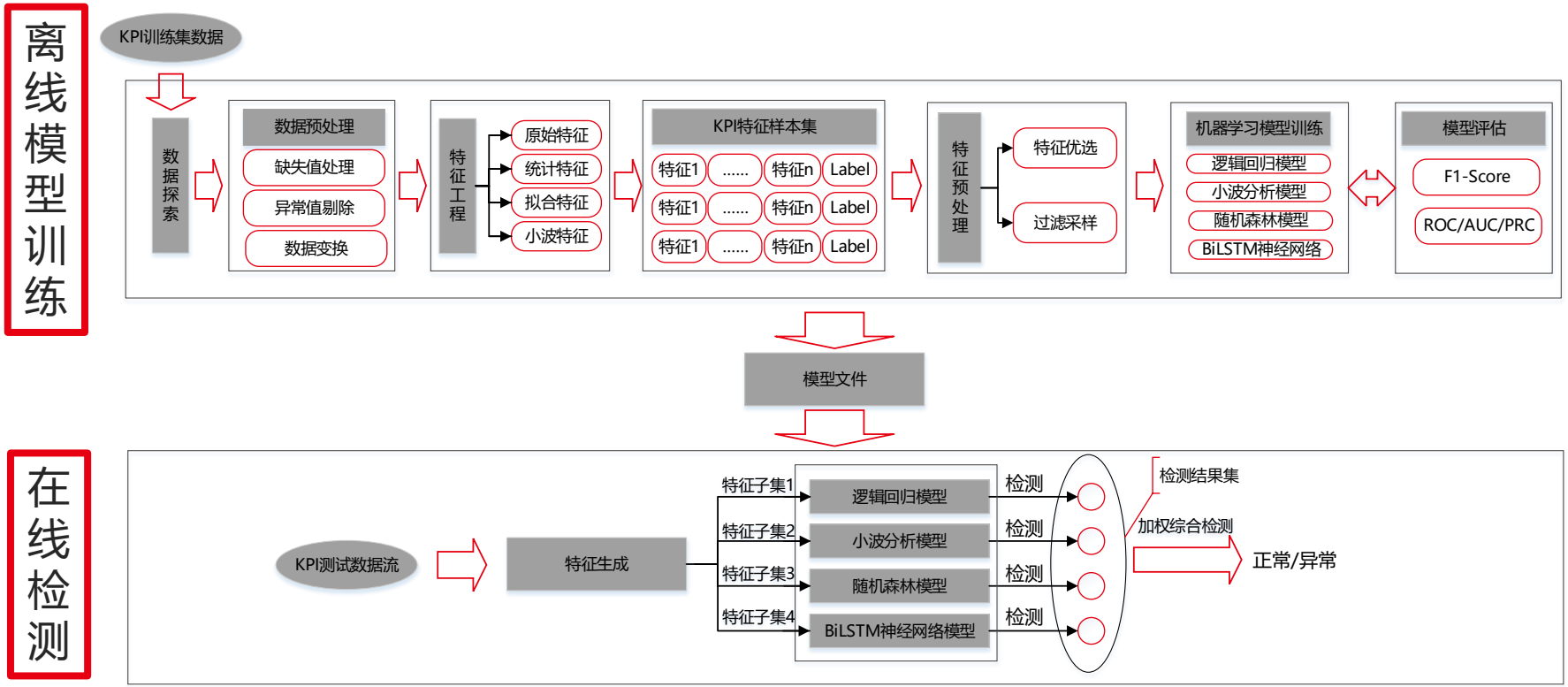
1. **异常发生的频率很低**：**负样本稀少**
2. **KPI异常类型：多种多样**
3. **KPI三种形态**：周期波动/稳定/不稳定型

➢ **负样本覆盖面不全，**测试集中部分异常类型没有出现在训练集中

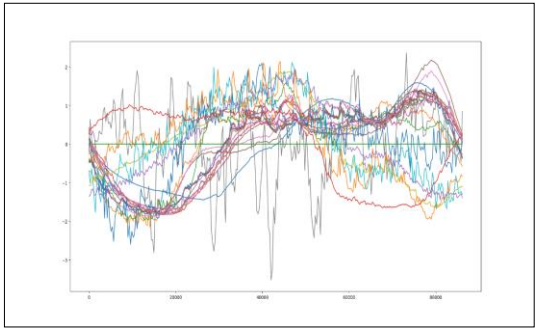
数据和**特征**决定了机器学习的上限

好的**算法**和**模型**可以逼近这个上限

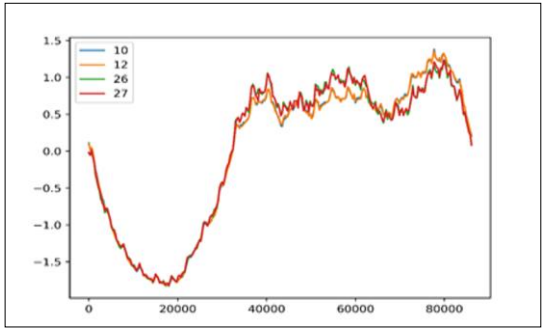
敏捷AI工程


离线模型训练
KPI训练集数据
数据探索
数据预处理
缺失值处理
异常值剔除
数据变换
特征工程
原始特征
统计特征
拟合特征
小波特征
KPI特征样本集
特征1 ..... 特征n Label
特征1 ..... 特征n Label
特征1 ..... 特征n Label
特征预处理
特征优选
过滤采样
机器学习模型训练
逻辑回归模型
小波分析模型
随机森林模型
BiLSTM神经网络
模型评估
F1-Score
ROC/AUC/PRC
模型文件
在线检测
KPI测试数据流
特征生成
特征子集1
特征子集2
特征子集3
特征子集4
逻辑回归模型
小波分析模型
随机森林模型
BiLSTM神经网络模型
检测
检测
检测
检测
检测结果集
加权综合检测
正常/异常

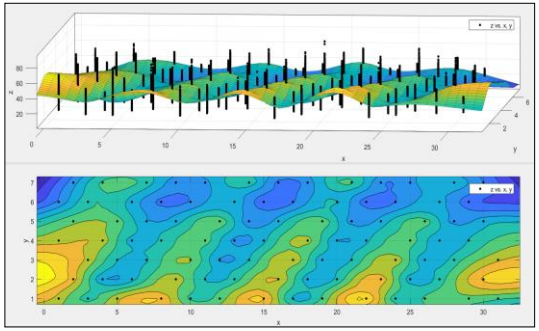
数据探索



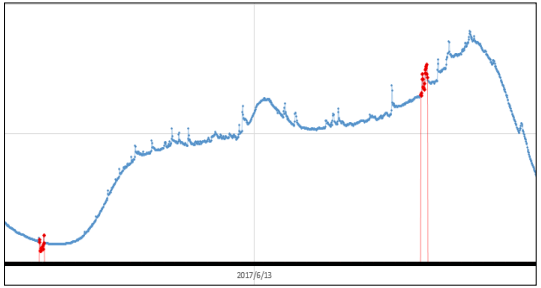
均值分析



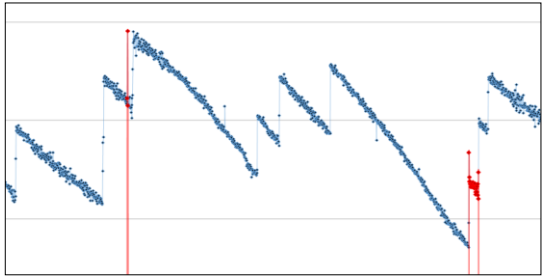
KPI间相似性分析



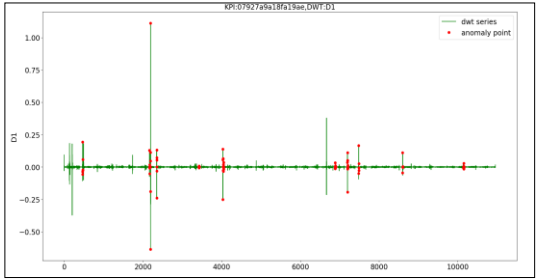
宏观分布分析



异常细节分析



异常细节分析

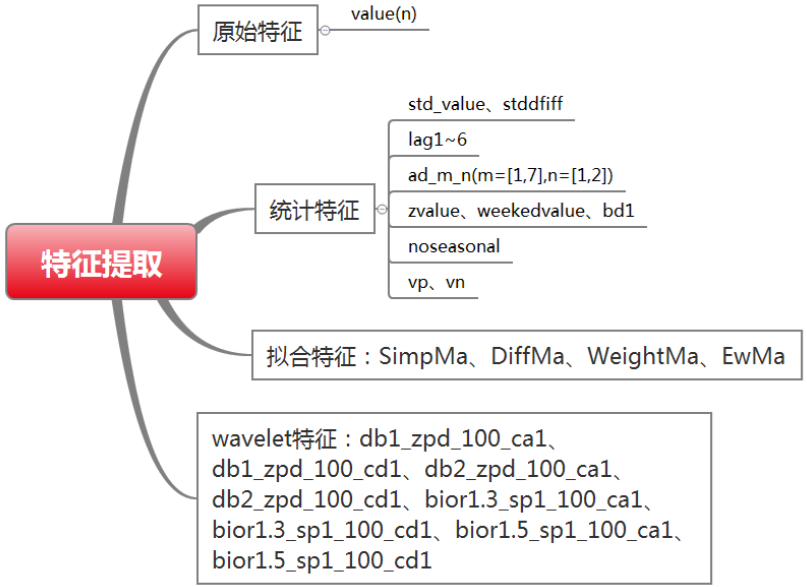


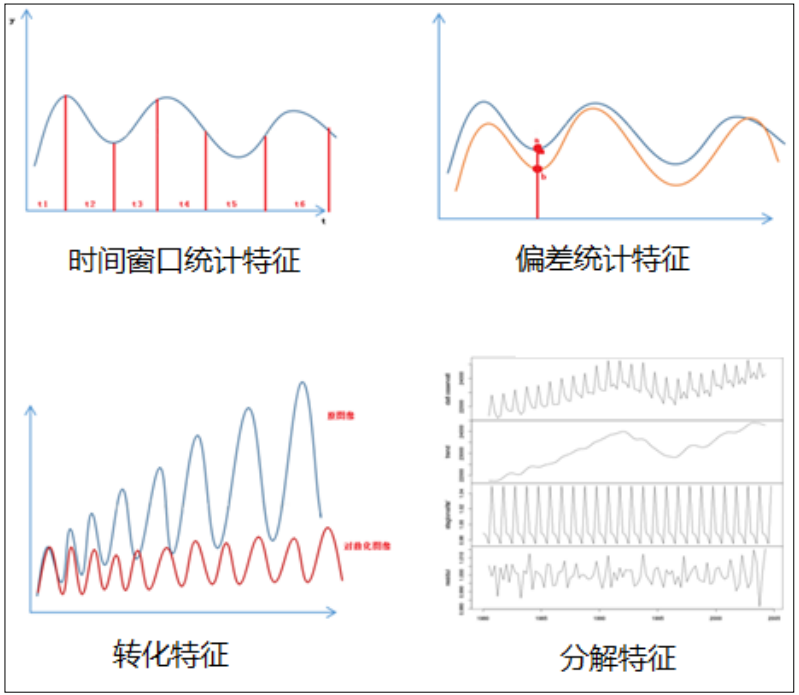
多尺度小波分解细节分析

特征提取

H3C
数字化解决方案领导者

提取了61个时序特征



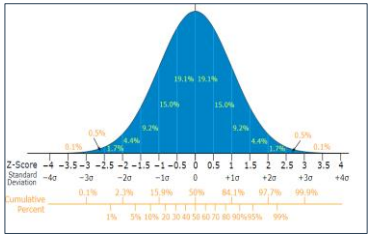


数据处理

数据预处理：

- 数据标准化：z-score、min-max
- 缺失值处理：均值补全
- 异常值处理：格拉布斯准则剔除异常点

$$x_{zscore} = \frac{x - \mu}{\sigma} \qquad x_{min-max} = \frac{x - min}{max - min}$$

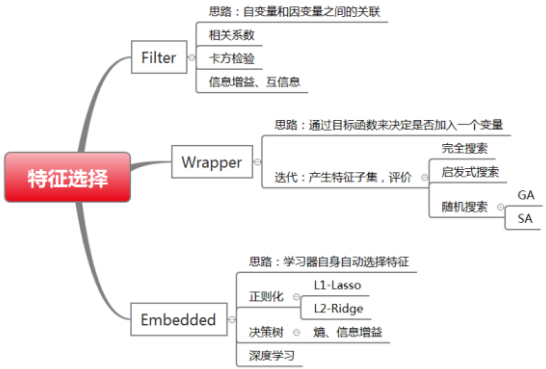


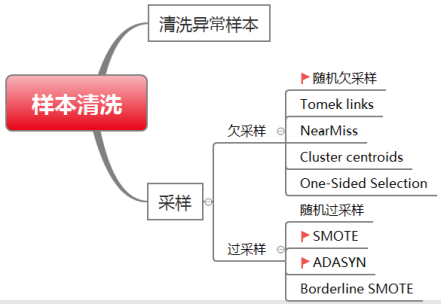
grubbs

$$G = \frac{\max_{i=1,\ldots,N} |Y_i - \bar{Y}|}{s}$$

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

特征/样本处理：





异常检测模型

01 逻辑回归模型

02 小波分析模型

03 随机森林模型

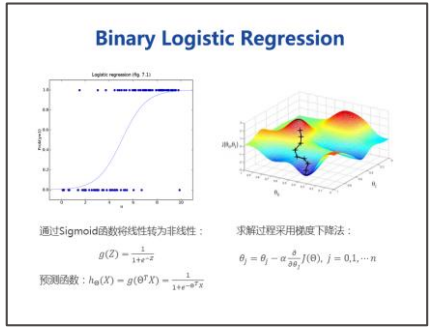
04 BiLSTM神经网络模型

基于逻辑回归的异常检测

H3C 数字化解决方案领导者

逻辑回归模型

- 能有效解决分类问题
- 简单高效，性能好
- 能快速进行模型训练



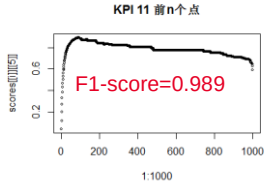
原理

逻辑回归的实现：

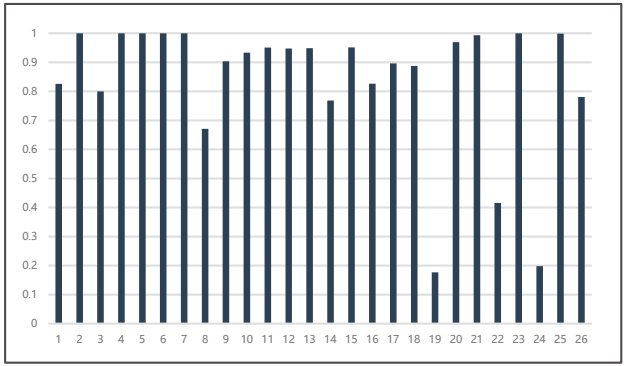
- 数据预处理
- 训练集数据STL分解
- 特征选取：vp,vn, ad_1_1~ad_7_1,ad_1_2~ad_7_2
- 选取训练集异常区间前7个点（对于少于7个点的异常区间则选取该区间的所有异常点）和所有正常点作为训练样本
- 模型训练，梯度下降求解参数
- 确定阈值
- 测试集标注

实现过程

- 训练样本的选择
- 阈值的确定



关键问题



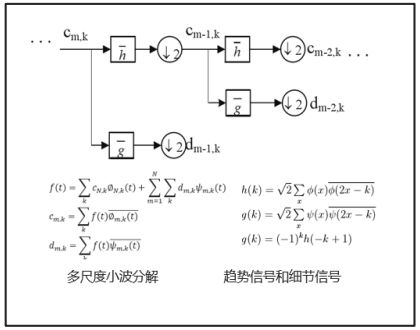
异常检测结果

基于小波分析的异常检测

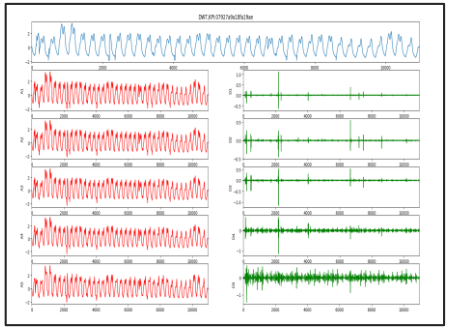
H3C
数字化解决方案领导者

时间序列多尺度小波分解

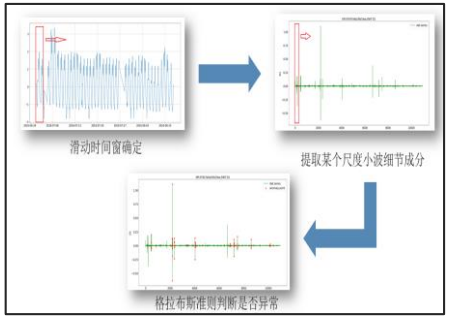
- 具有良好的时频和局部特征
- 时间序列=趋势成分+细节成分
- 细节成分可被用于异常检测



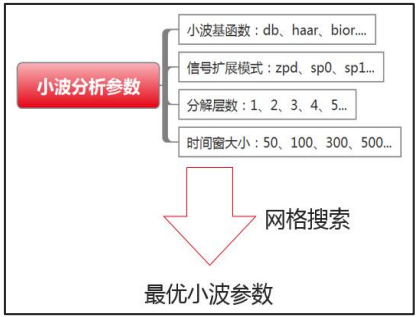
多尺度小波分解原理



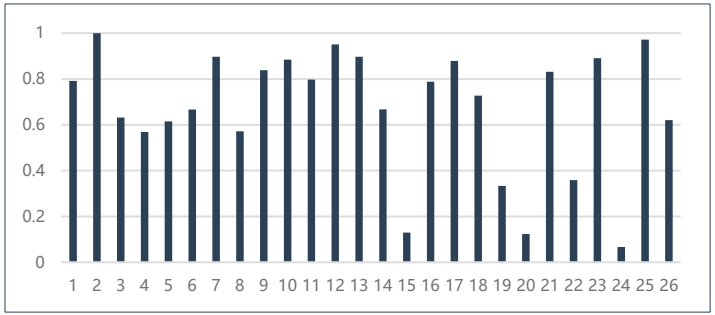
多尺度小波分解示例



实现过程



参数优化



异常检测结果

基于随机森林的异常检测

H3C 数字化解决方案领导者

随机森林

- ✓ 能处理高维度数据
- ✓ 对数据要求低，鲁棒性强
- ✓ 是一种有效的异常检测模型

参数优化

- ➢ 树的数目：n_estimators
- ➢ 最大深度：max_depth
- ➢ 最大特征数：max_features
- ➢ 叶子节点最小样本数：

  min_sample_leaf

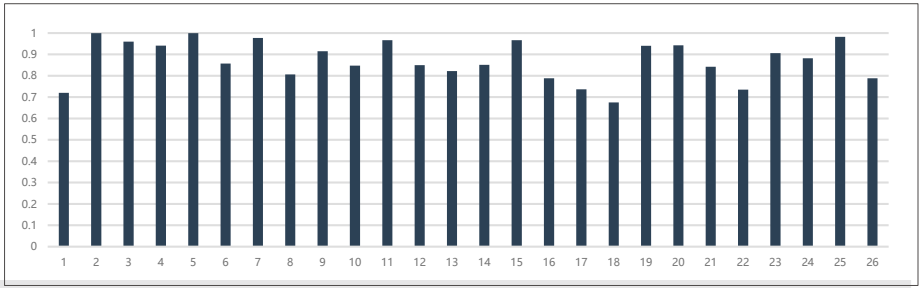
  ……

网格搜索优化

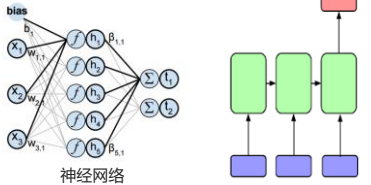
实现代码：

```
def Rf_train(kpiid,dfTrain, dfTrainFeature, selectedFeatures,
                    persist_path,ratio=1.0, anomalyClean=True):
    selectedFeatures=np.array(selectedFeatures)
    dfAnom=dfTrainFeature[dfTrainFeature['label']==1]
    dfNormal=dfTrainFeature[dfTrainFeature['label']==0].sample(frac=ratio)
    df=dfAnom.append(dfNormal)
    df=shuffle(df)
    X=df[selectedFeatures].values
    y=df['label'].values
    X_all=dfTrainFeature[selectedFeatures].values
    rf = RandomForestClassifier(random_state=1377)
    rf.fit(X, y)
    y_pred=rf.predict(X_all)
```

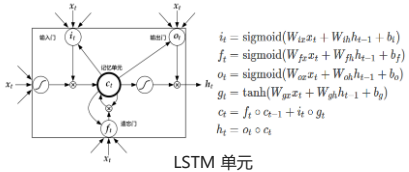
训练集得分：



基于BiLSTM 循环神经网络的异常检测

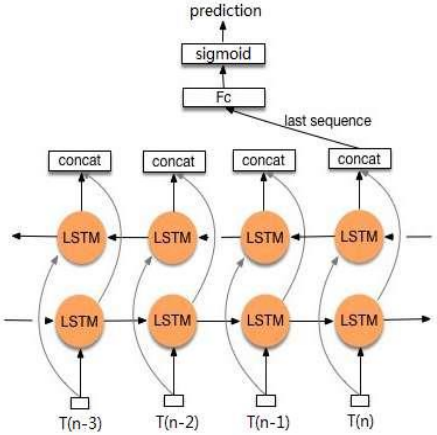


神经网络



$$i_t = \text{sigmoid}(W_{ix}x_t + W_{ih}h_{t-1} + b_i)$$
$$f_t = \text{sigmoid}(W_{fx}x_t + W_{fh}h_{t-1} + b_f)$$
$$o_t = \text{sigmoid}(W_{ox}x_t + W_{oh}h_{t-1} + b_o)$$
$$g_t = \tanh(W_{gx}x_t + W_{gh}h_{t-1} + b_g)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ g_t$$
$$h_t = o_t \circ c_t$$

LSTM 单元



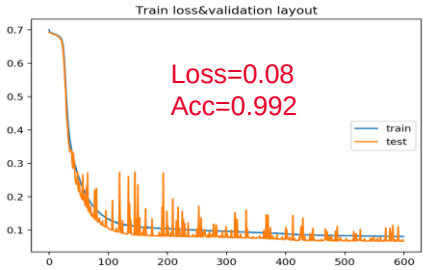
- 神经网络学习能力强
- LSTM是RNN优秀变种
- 适合处理时序数据相关问题
- 双向结构BiLSTM性能更好

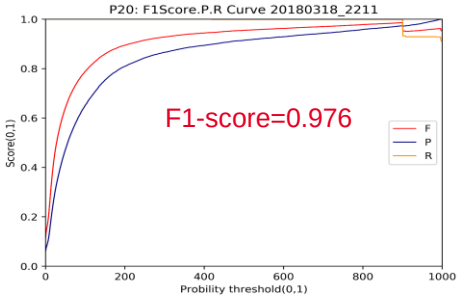
使用BiLSTM构建
多对一时序预测模型
直接用于异常检测

难点：模型训练

主要超参数：

- 学习率 η
- 正则化参数 λ
- 神经网络的层数 L
- 每一个隐层中神经元的个数
- 学习的回合数Epoch
- 小批量数据 batch size
- 输出神经元的编码方式
- 代价函数的选择
- 权重初始化的方法
- 神经元激活函数的种类
- 参加训练模型的数据





综合模型异常检测

H3C
数字化解决方案领导者

1. 各模型进行单独异常检测

2. 加权投票方法进行综合判断

$$P = \frac{\alpha_{logit}P_{logit} + \alpha_{wavelet}P_{wavelet} + \alpha_{rf}P_{rf} + \alpha_{lstm}P_{lstm}}{\alpha_{logit} + \alpha_{wavelet} + \alpha_{rf} + \alpha_{lstm}}$$

α_i：为各模型权重，取值0~1

P_i：各模型异常检测结果，取值0或1

P：综合预测结果，如果P<0.5，则预测正常，否则异常

权重处理：

情形1：存在得分大于0.5的模型的条件下，若某个模型的训练集得分小于0.5，则对应的的权重置为0

情形2：若所有模型的得分小于0.5，则取4个模型中得分最高的作为预测模型，对应的其他模型的权重置为0

测试集成绩分析：

H3C
数字化解决方案领导者

决赛成绩 (综合模型)

队伍排名	队伍名称	队伍分数
1	LogicMonitor-AI	0.79567
2	D.I(H3C)	0.771397
3	ICA128	0.73492
4	火眼金睛	0.721988
5	烧脑特工队	0.645889

预赛成绩 (BiLSTM)

每页显示 10 记录

队伍排名	队伍名字	队伍分数
1	D.I(H3C)	0.821641752084
2	ICA128	0.801035034811
3	LogicMonitor-AI	0.7895230619
4	火眼金睛	0.787489192136
5	烧脑特工队	0.783470506235

比赛阶段	测试集表现	策 略	模 型
决 赛	**0.7714 (综合)**	稳妥方案	综合模型
预 赛	**0.8216 (BiLSTM)**	积极方案	单一模型

实际表现： 分数下降

主要原因： 策略选择不当

改进方案：

- 继续优化单模型性能
- 改进多模型融合方式

总结

H3C
数字化解决方案领导者

模型的优点：

- 新颖性：使用BiLSTM循环神经网络，直接进行异常检测；
  采用融合模式，可动态集成其他检测模型。
- 智能化：自适应、自主学习、异常检测高效、准确
- 健壮性：采用融合模式，算法稳定可靠，风险度低。
- 普适性：可用于多种类型的KPI的异常检测

敏捷AI工程：

- 团队分工协作，充分交流
- 快速迭代、快速验证、快速改进

展望

H3C
数字化解决方案领导者

1、优化改进：

- 特征优选
- 模型优化
- 集成优化

2、模型应用

- 比赛成果产品化，应用于更多实际项目
- 跨领域应用：金融、安全领域等

Q&A

我们深知运维的过去和现在

更致力于变革运维的未来

以梦想改变世界

紫光集团.新华三

www.h3c.com