智能运维KPI异常检测决赛答辩

ICA128

2018. 05. 19

团队介绍

ICA128　中国工程物理研究院计算机应用研究所128室

褚新坤（领队）　　算法设计与实现

chuxk@pku. edu. cn　　2017毕业于北京大学　博士　高能物理数据分析

田志宇　　算法实现

2014毕业于西北工业大学　硕士　飞行器试验数据分析

张浩　　数据分析

2010毕业于国防科技大学　博士　应用数学

目录

一. 赛题理解

基于互联网公司真实运维环境的KPI异常检测

1. 异常检测算法的普适性

- 适用于不同特征类型KPI、无需手动调参 → 集成算法
- 异常种类多样而样本稀少 → “正常”模式学习

2. 异常检测的实时性

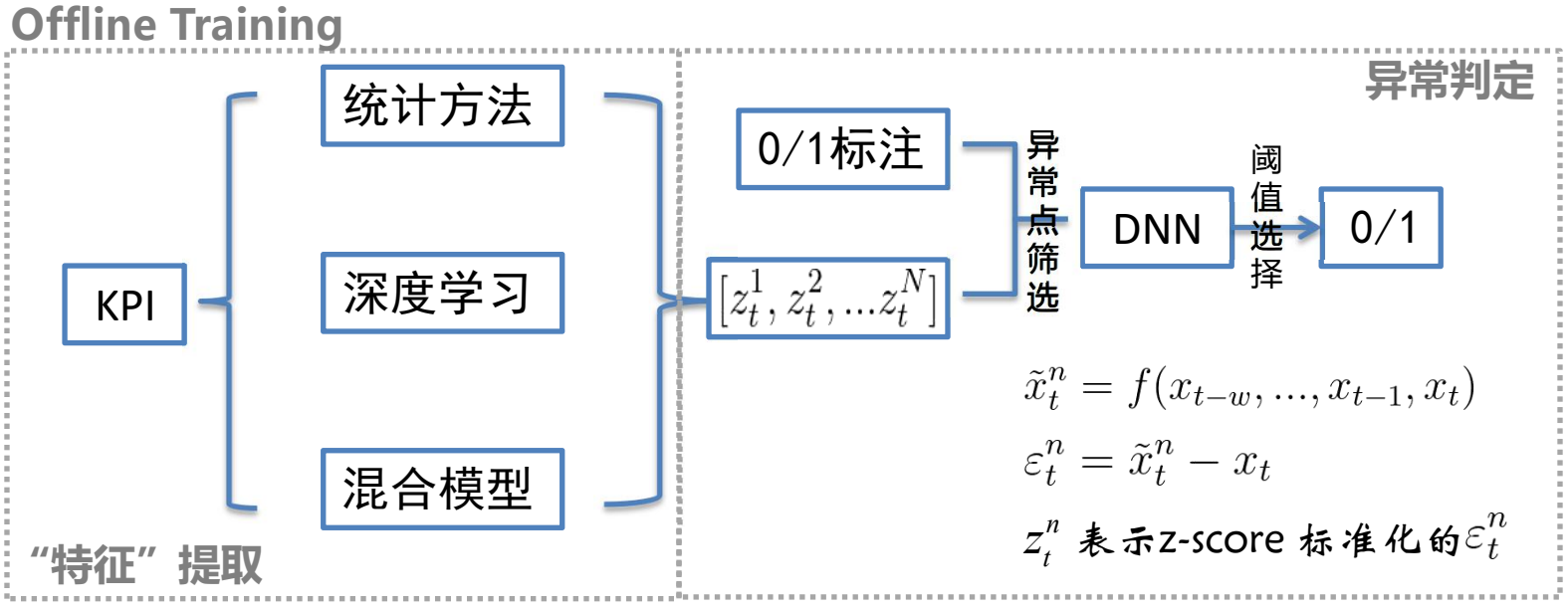
- 在线检测(delay=7)、不能利用未来数据 → 基于单点判别
- 算法的时间性能

二. 数据预处理

KPI ID	缺失值%	缺失区间长度 min, max, median	异常值%	异常区间长度 min, max, median
02e9	2.5%	1, 1658, 8	8.0%	1, 840, 103
8c89	3.1%	1, 1331, 1	2.0%	1, 133, 10
c58b	2.4%	1, 1774, 2	0.1%	1, 10, 4
affb	2.2%	1, 312, 1	0.4%	1, 154, 4
……				

1. 按ID分开数据，分别训练
2. 对缺失值按时间点补齐后用一阶线性差值填充
3. 对参数需要从数据中拟合的模型，用线性差值代替异常点用于模型训练

三. 建模 -- 算法框架



Ref. Dapeng Liu, et.al. Opprentice: Towards Practical and Automatic Anomaly Detection Through Machine Learning”. ACM IMC 2015

三. 建模 --"特征"提取

统计方法

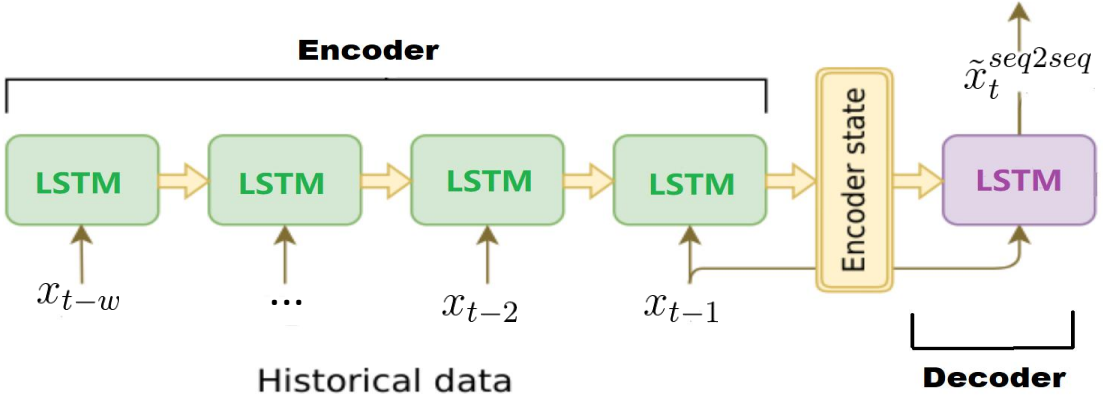
Model	Sampling parameters
MA/ MA_Diff	window = 5, 10, 20
EWM	α = 0.1, 0.3, 0.5, 0.7, 0.9
2nd.Exp.Smooth	α, β = {(0.2,0.4), (0.4,0.6), (0.6,0.4), (0.8,0.2)}
LOWESS/LOESS	span = 10
Diff	last-slot, last_day
Wavelet	db4, coif4, sym8, dmey, bio2.8, rbio2.8, haar
Holt-Winters	α, β, γ fitted from data/2days
AR/ARIMA	fitted from data

三. 建模 --"特征"提取

深度学习

LSTM、Seq2Seq

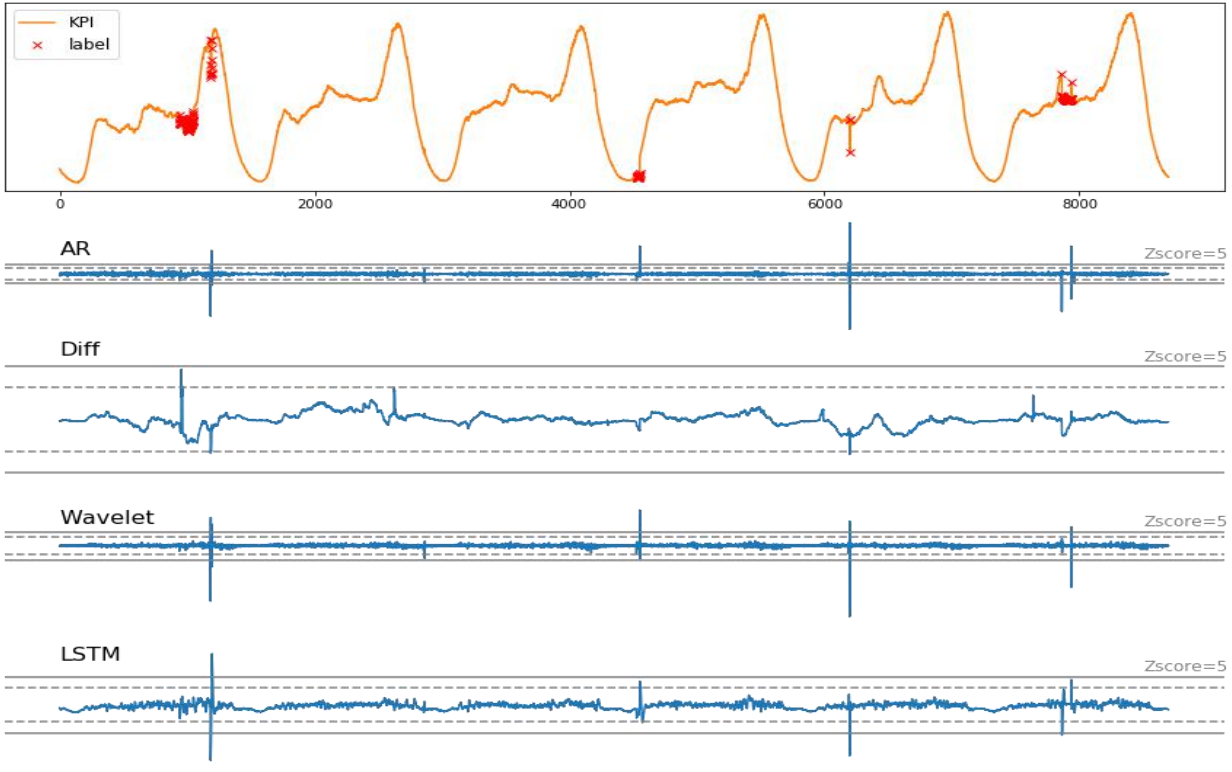
Seq2Seq结构示意图



混合模型

AR + LSTM	$x_t \xrightarrow{AR} \tilde{x}_t^{AR} - x_t \xrightarrow{LSTM} \tilde{x}_t^{AR+LSTM}$
STD(SeasonalTrend Decomposition) +	$x_t \xrightarrow{STD} x_t - ssn. \xrightarrow{\text{stat. models}} \tilde{x}_t^{STD+}$

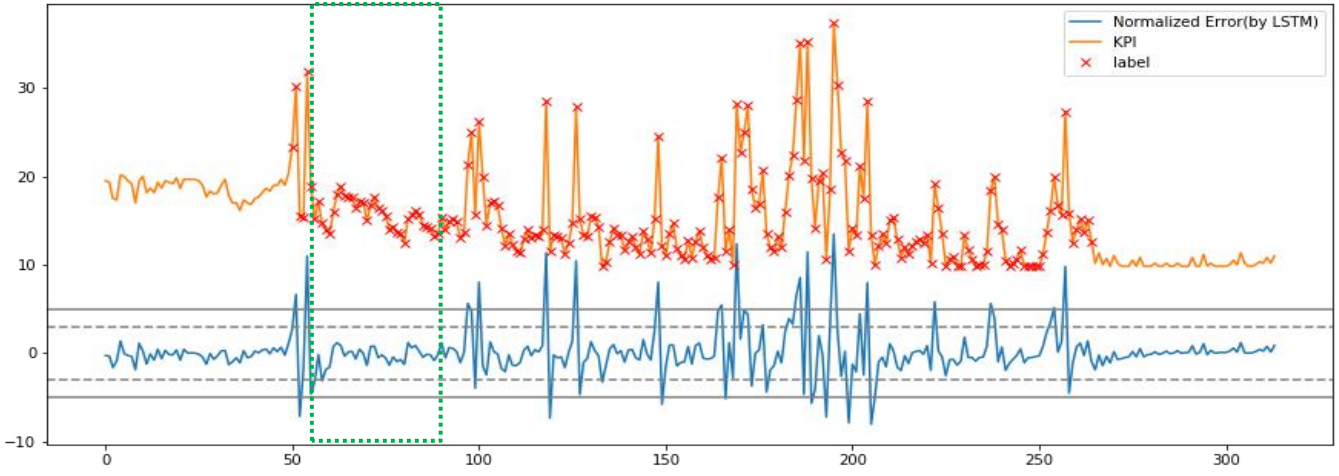
三. 建模 --"特征"提取



部分特征提取算法计算的 z_t^n 与标注了异常的KPI对比

三. 建模 -- 异常判定

异常点筛选

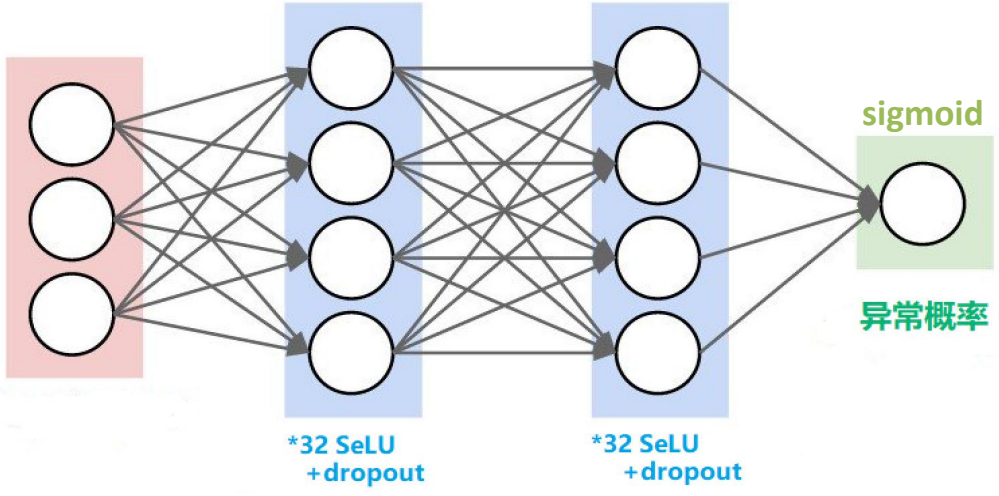


- 长异常区间只取前8个时间点(delay=7)
- $\exists n, \quad z_t^n > 3$ 排除”误”标注的点
- 异常点上采样(>2%)

三. 建模 -- 异常判定

DNN

$[z_t^1, z_t^2, ...z_t^N]$



阈值确定

精确度(P)和回召率(R)的权衡

选取阈值使调整后算法输出的 F-score {2PR/(P+R)}最大

四. 总结讨论

预赛 F-score 0.801 2nd

决赛 F-score 0.735 3rd

决赛基于7*24h时间限制

- 降低了部分特征提取算法的复杂度
- 取消了NAR、Seq2Seq等模型

改进方向

- 超参数的优化；
- 动态阈值确定；
- 相似度高的KPI样本一起训练；
- 特征提取算法的并行，更多差异化的特征提取算法(如VAE)

感谢主办方的辛勤付出！

感谢在场聆听的专家评委！

感谢比赛中共同进步的朋友！