织云Metis时间序列异常检测全方位解析

Waywang（汪华）

腾讯社交网络运营部——Metis智能运维团队

Geekbang. 极客邦科技 | 极客时间

SPEAKER

INTRODUCE

汪华　高级工程师

- 云管理解决方案开发
- 手机QQ、QQ会员等业务运维
- 运维自动化建设
- Metis智能运维建设

Tencent 腾讯　QQ　QQ会员　QQ音乐　QQ空间

SPEAKER

ArchSummit 2018`ShenZhen

TABLE OF

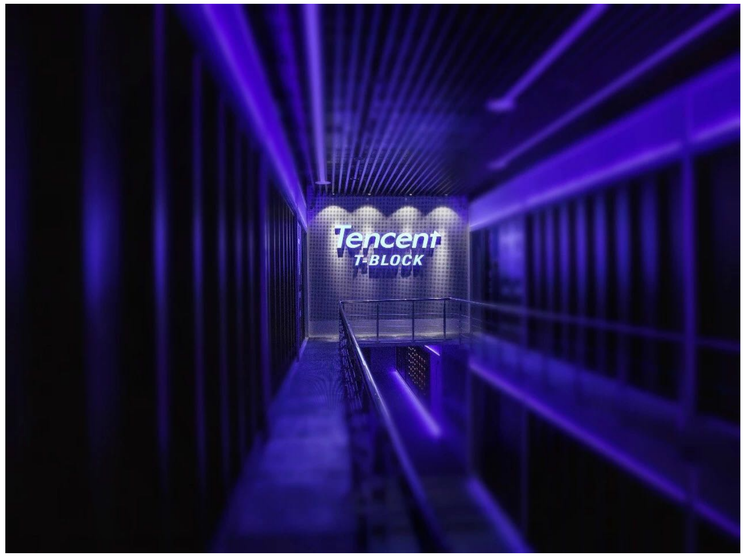
CONTENTS 大纲

- 传统时序监控的问题与新思路
- 检测算法原理与应用
- 特征工程与打标工程
- 样本库建设与管理
- Metis概述（智能运维应用实践）

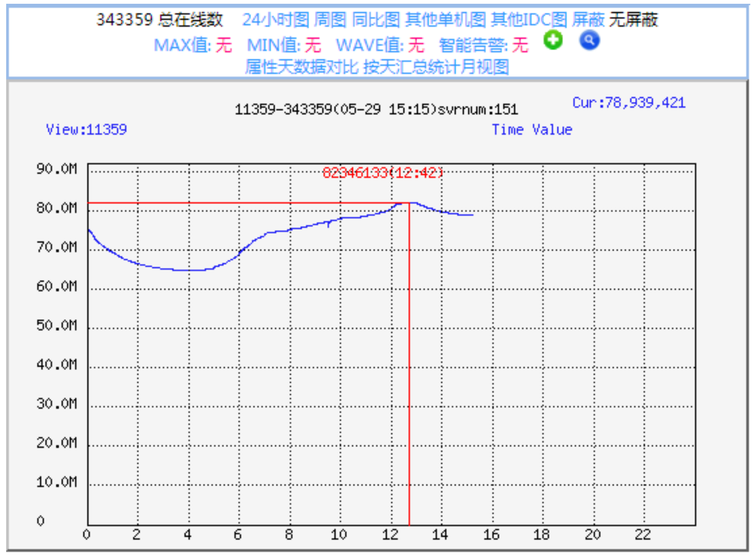
业务规模

- 轻微的异常就会影响到大量的外网用户

在线2.8亿
月活8.05亿



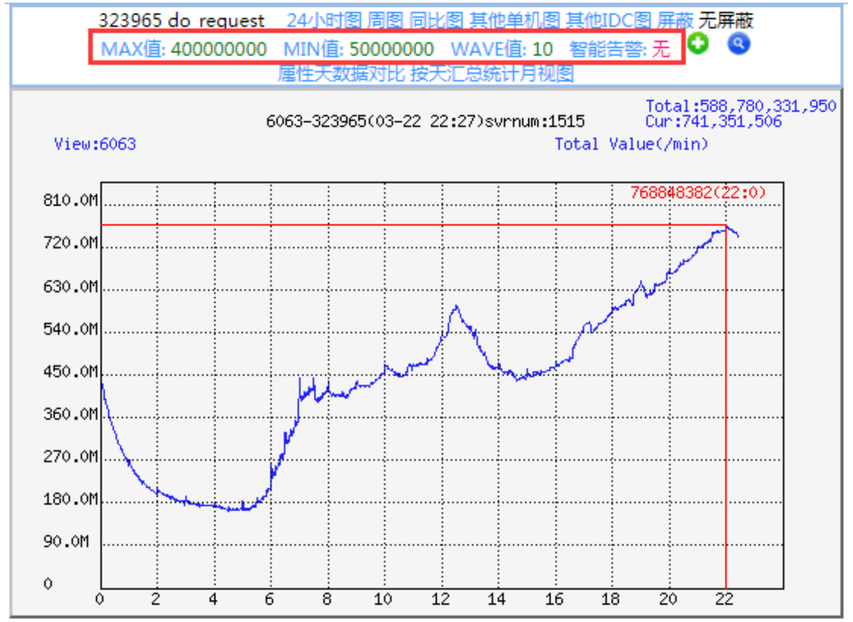
SNG服务器20w+

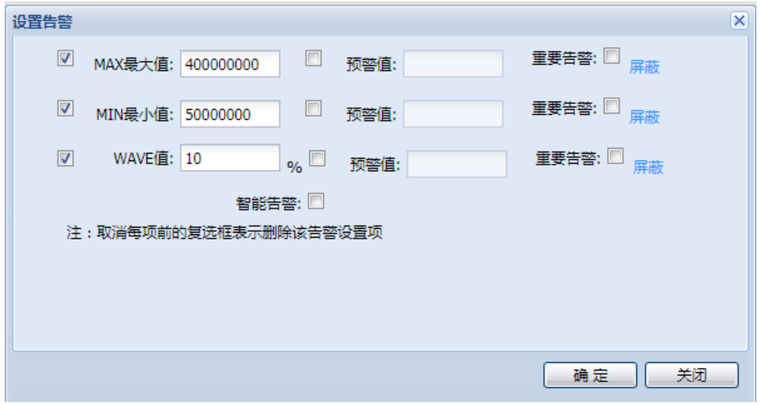


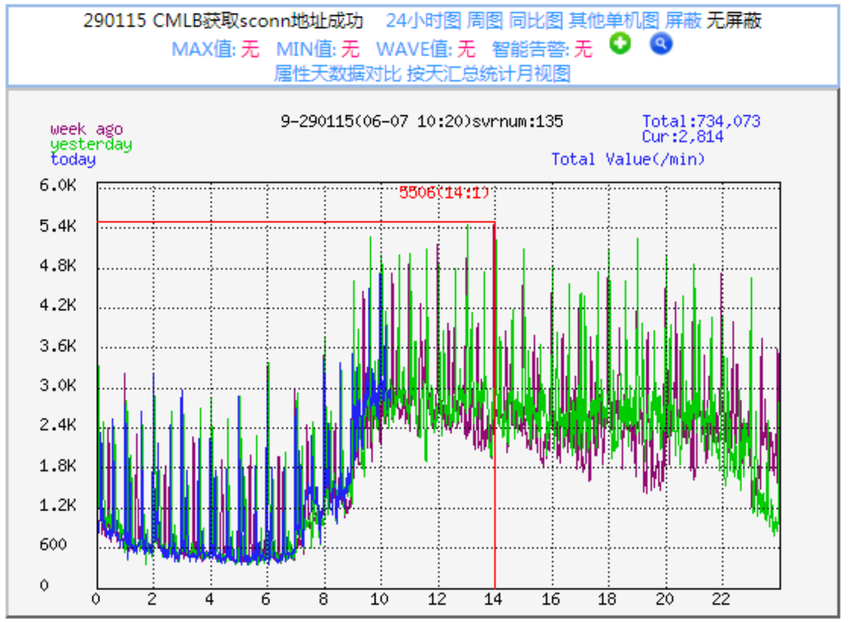
社交类指标240w+

传统监控与新思路

- 随着业务发展，传统监控呈现出的一些问题







准确率低

维护成本高

形态各异

传统监控与新思路

算法和机器学习的新思路是否可应用？

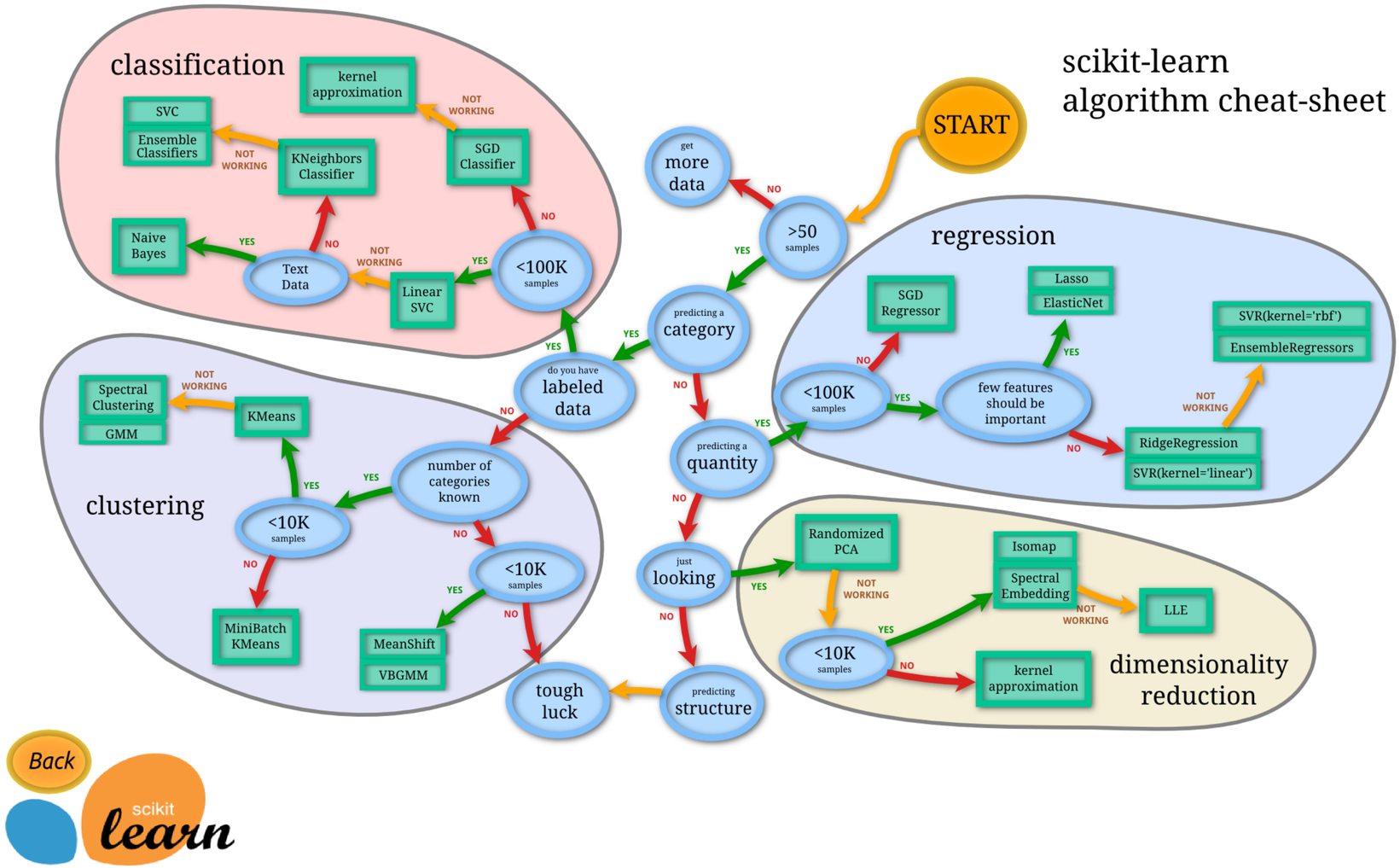


TABLE OF

CONTENTS 大纲

- 传统时序监控的问题与新思路
- 检测算法原理与应用
- 特征工程与打标工程
- 样本库建设与管理
- Metis概述（智能运维应用实践）

从大量输入中总结出准确预测的规律（模型）

$$y = f(x) = ax + d$$

$$y = f(x) = ax^2 + bx + d$$

$$y = f(x_1, x_2, x_3) = ax_1 + bx_2 + cx_3 + d$$

$$a_1^{(2)} = f(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)})$$
$$a_2^{(2)} = f(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)})$$
$$a_3^{(2)} = f(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)})$$

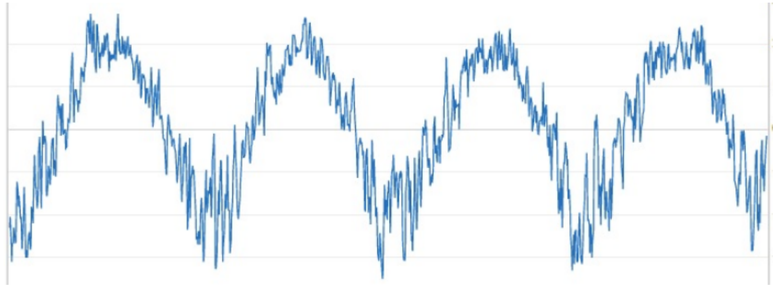
$$h_{w,b}(x) = a_1^{(3)} = f(W_{11}^{(2)}a_1^2 + W_{12}^{(2)}a_2^2 + W_{13}^{(2)}a_3^2 + b_1^{(2)})$$

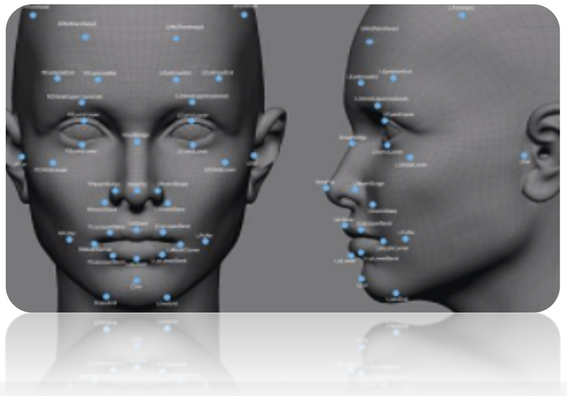
x, x_1, x_2, x_3是我们的输入

y是期望的输出

a,b,c,d （W，b）是我们要求解的参数

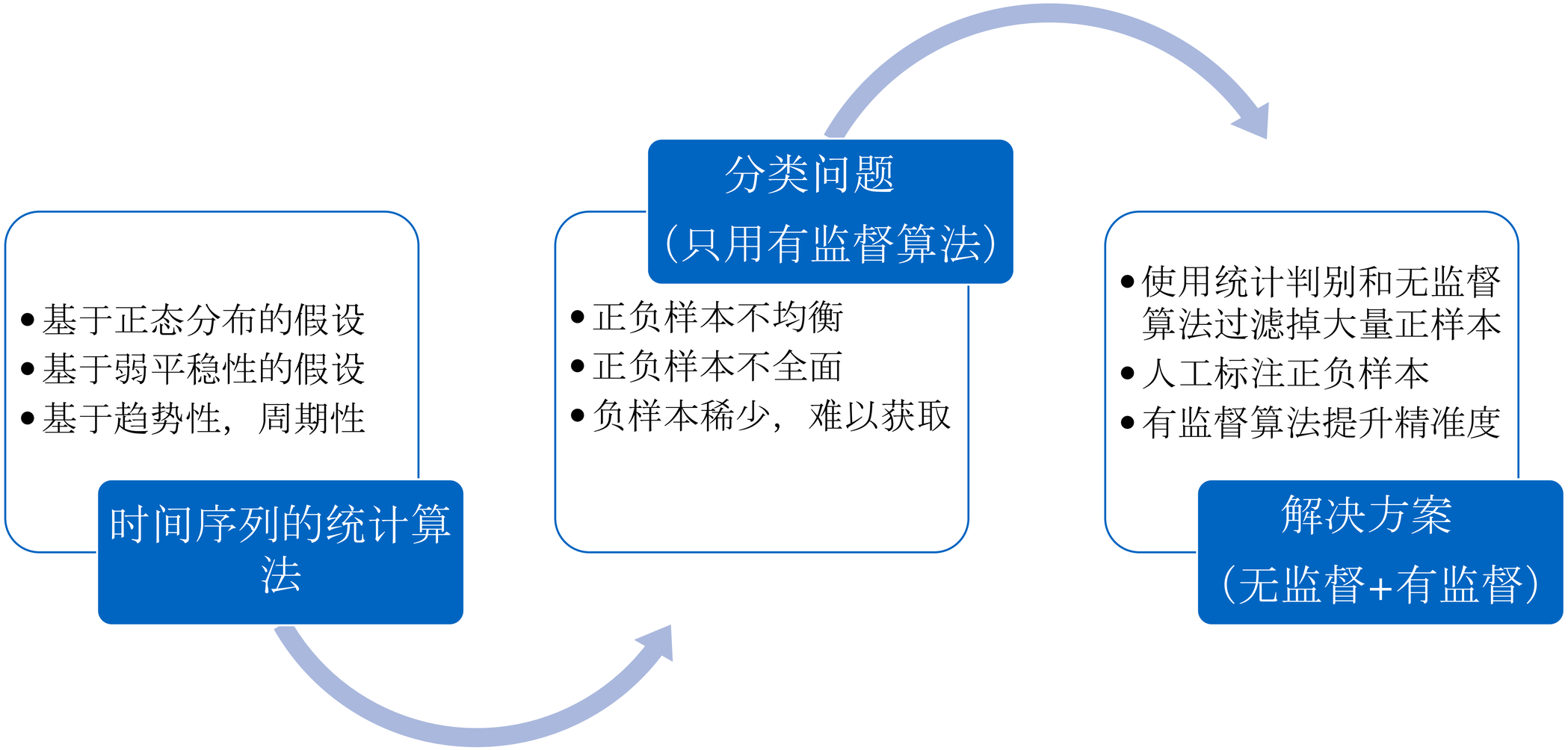
f是转换函数（公式、函数、算法、模型…）也就是我们要找的“规律”





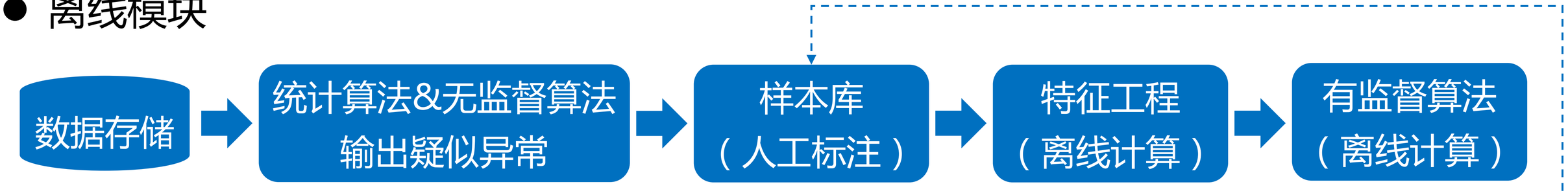
数值型预测
0/1型预测
概率型预测等

技术路线演进


时间序列的统计算法
• 基于正态分布的假设
• 基于弱平稳性的假设
• 基于趋势性，周期性
分类问题
（只用有监督算法）
• 正负样本不均衡
• 正负样本不全面
• 负样本稀少，难以获取
解决方案
（无监督+有监督）
• 使用统计判别和无监督算法过滤掉大量正样本
• 人工标注正负样本
• 有监督算法提升精准度

技术框架

- 离线模块



人工审核

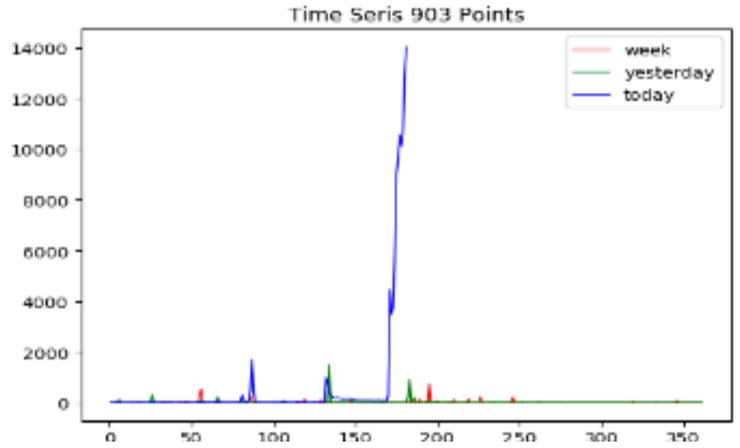
- 在线模块

- ABTest模块

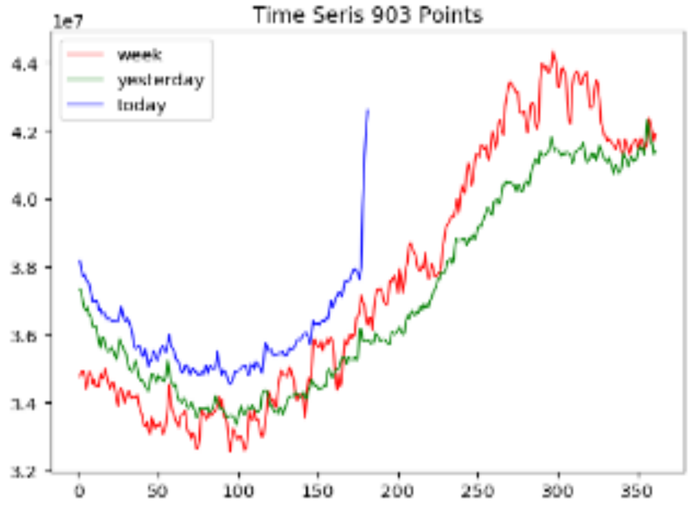
第一层：统计判别算法

- 3sigma算法与控制图算法的优缺点

√



×



数据提取

- 以当前时刻为标准
- 七天前后三小时 + 昨天前后三小时 + 今天前三小时

Grubbs

- 3sigma原理

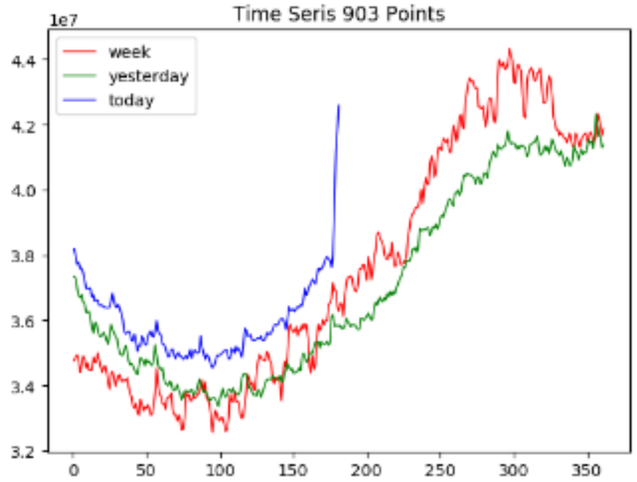
控制图

- 移动平均算法
- 指数移动平均算法

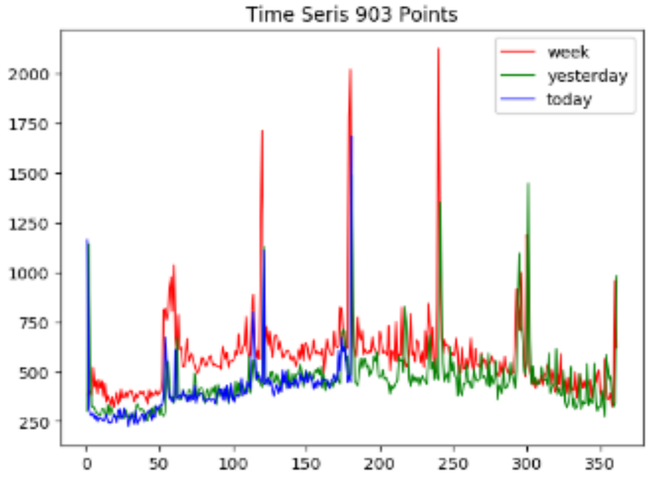
第一层：无监督算法

- 无监督学习算法的优缺点

√



×



数据提取

- 以当前时刻为标准
- 七天前后三小时 + 昨天前后三小时 + 今天前三小时

孤立森林

- 可以从多维特征中寻找异常点

SVM

- 使用超平面的思想来进行异常/正常的区分

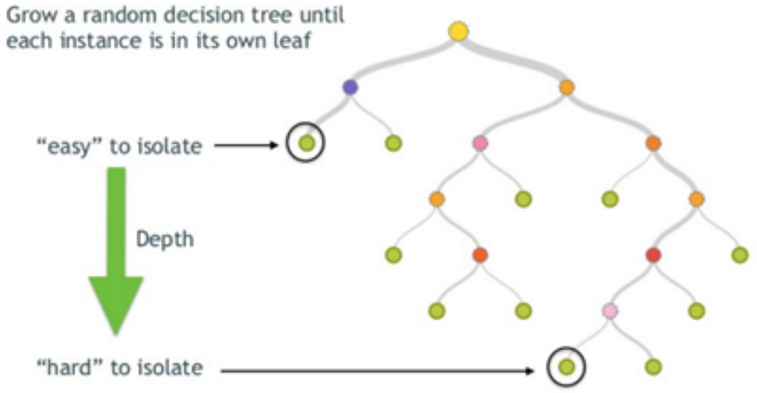
RNN

- 使用神经网络的误差来进行异常判断

第一层：无监督算法

Isolation Forest

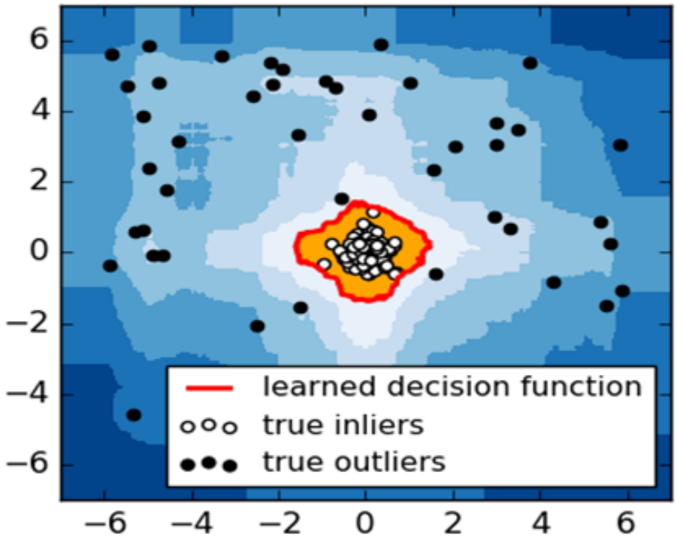
- 属于无监督算法
- 集成学习的思想
- 适用于连续数据的异常检测
- 通过多颗 iTree 形成森林来判断是否异常



Algorithm 1 $iTree(X, e, h)$

Input: X - input data; e - current height; h - height limit.
Output: an $iTree$.

1: **if** $e \geqslant h$ OR $|X| \leqslant 1$ **then**
2: **return** $exNode\{Size \leftarrow |X|\}$;
3: **else**
4: Randomly select an attribute q;
5: Randomly select a split point p between min and max values of attribute q in X;
6: $X_l \leftarrow filter(X, q < p)$, $X_r \leftarrow filter(X, q \geqslant p)$;
7: **return** $inNode\{ Left \leftarrow iTree(X_l, e+1, h)$,
$Right \leftarrow iTree(X_r, e+1, h)$,
$SplitAttr \leftarrow q, SplitValue \leftarrow p\}$;
8: **end if**



第一层：无监督算法

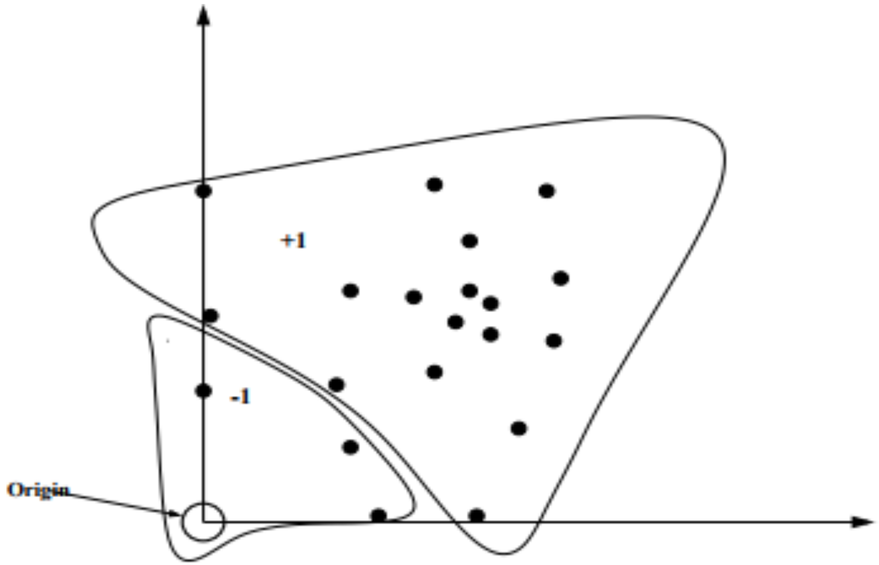
One Class SVM

- 属于无监督算法
- 使用了超平面的思想
- 适用于连续数据的异常检测
- 适用于对样本进行一定比例的筛选
- 寻找高维平面区分正常点与异常点

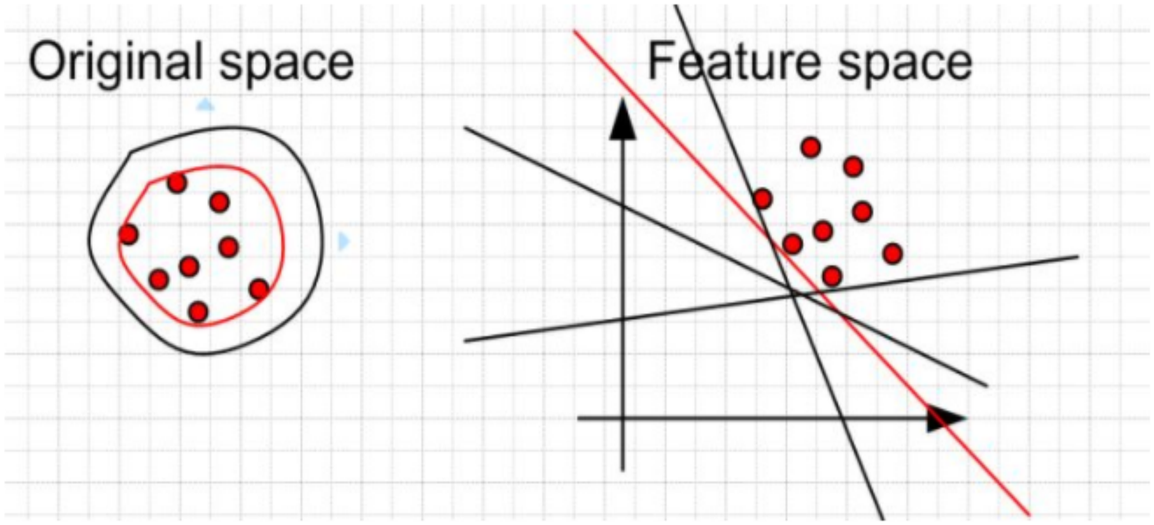
$$min\frac{1}{2}\|w\|^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho$$

subject to

$$(w \cdot \Phi(x_i)) \geq \rho - \xi_i \quad i = 1, 2, ..., l \quad \xi_i \geq 0$$



One-Class SVM Classifier. The origin is the only original member of the second class.

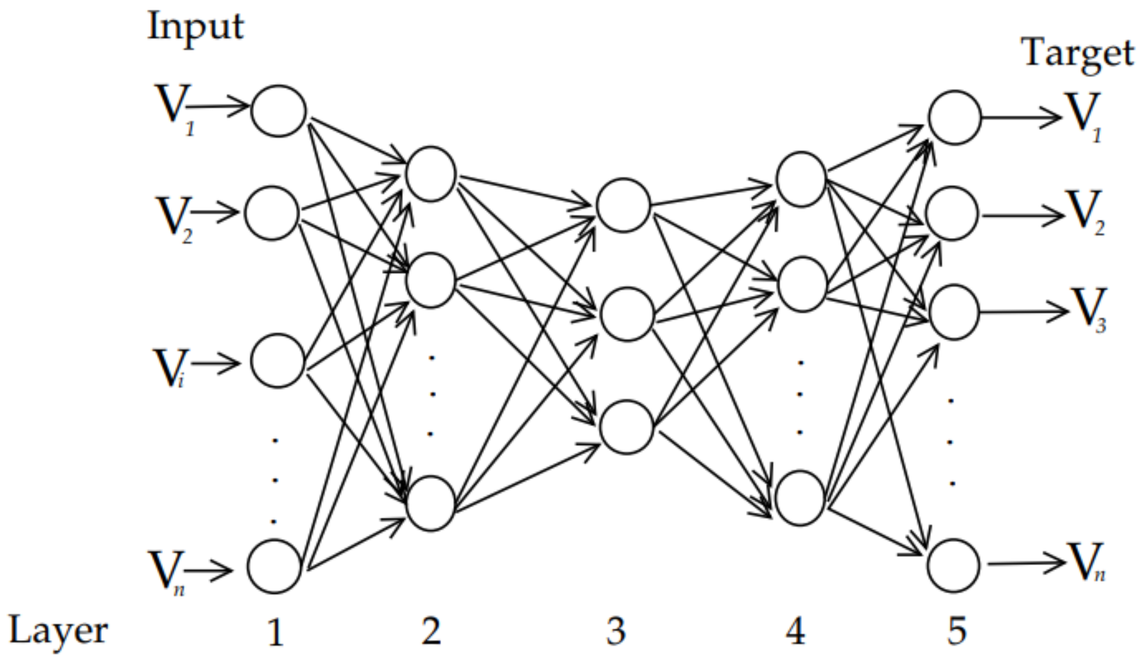


第一层：无监督算法

Replicator Neural Network

- 属于无监督算法
- 需要构造必要的特征
- 使用了神经网络的思想
- 适用于连续数据的异常检测
- 寻找神经网络的误差来区分正常点与异常点

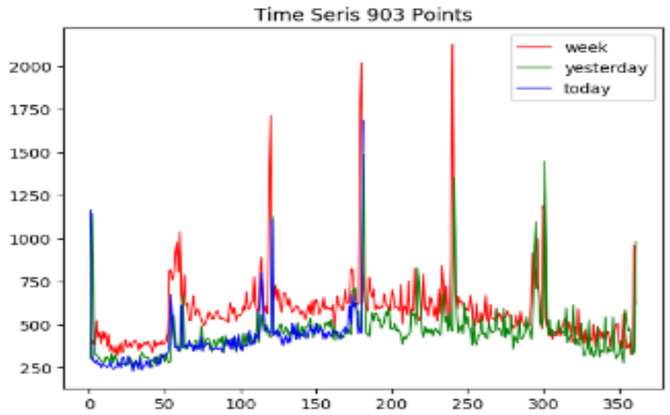
$$OF_i = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - o_{ij})^2$$



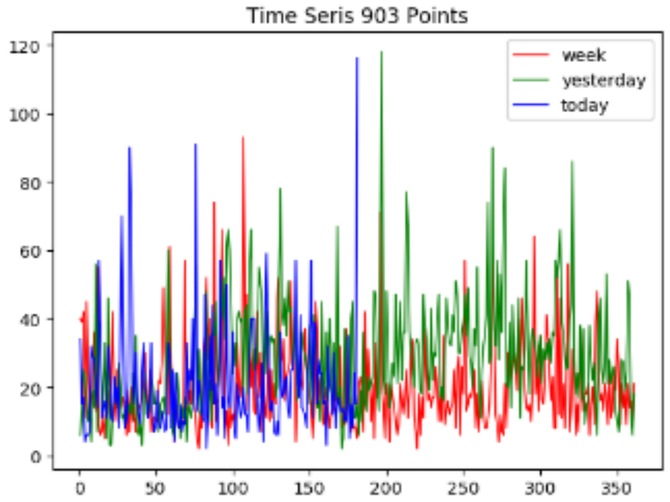
第二层：有监督算法

- 有监督算法能解决的问题

√

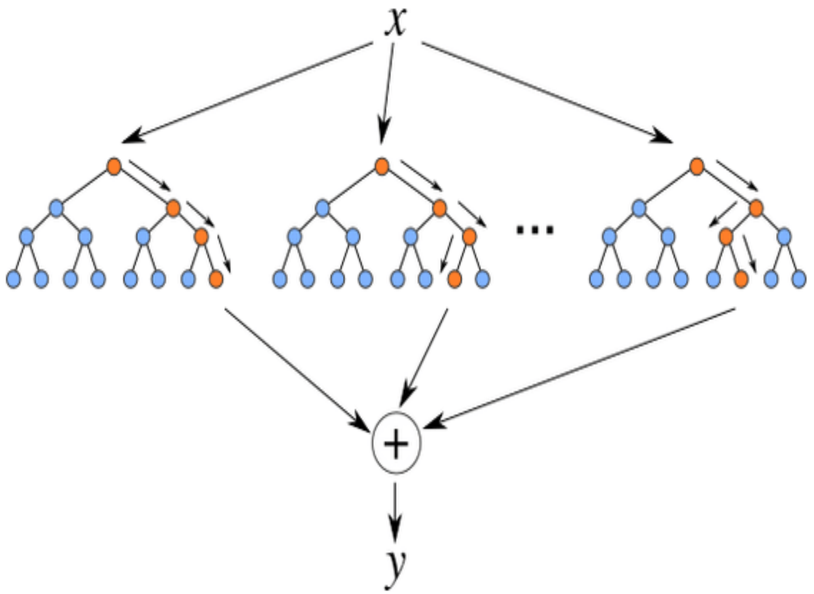


√



有监督算法

- Linear Regression/Logistic Regression
- Decision Tree/Naïve Bayes
- Random Forest/GBDT/xgboost



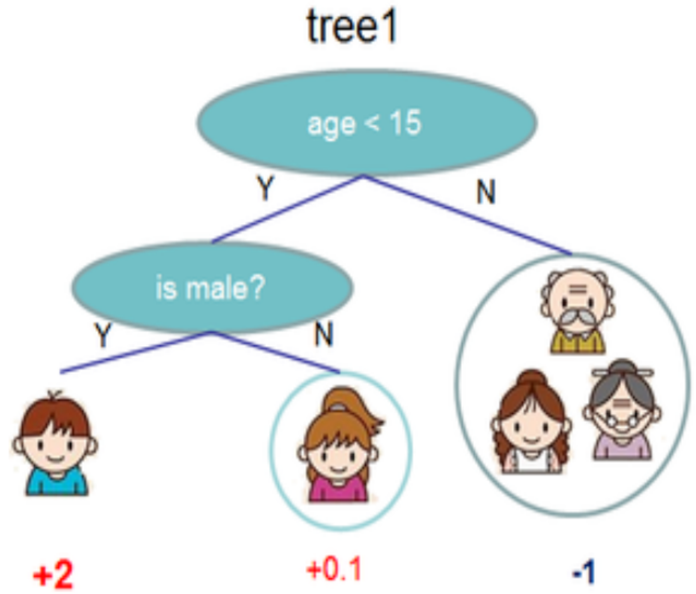


TABLE OF

CONTENTS 大纲

- 传统时序监控的问题与新思路
- 检测算法原理与应用
- 特征工程与打标工程
- 样本库建设与管理
- Metis概述（智能运维应用实践）

特征工程

统计特征

- 最大值，最小值，值域
- 最小值位置、最大值位置
- 均值，中位数
- 平方和，重复值
- 方差，偏度，峰度
- 同比，环比，周期性
- 自相关系数，变异系数

拟合特征

- 移动平均算法
- 带权重的移动平均算法
- 指数移动平均算法
- 二次指数移动平均算法
- 三次指数移动平均算法
- 奇异值分解算法
- 自回归算法
- 深度学习算法

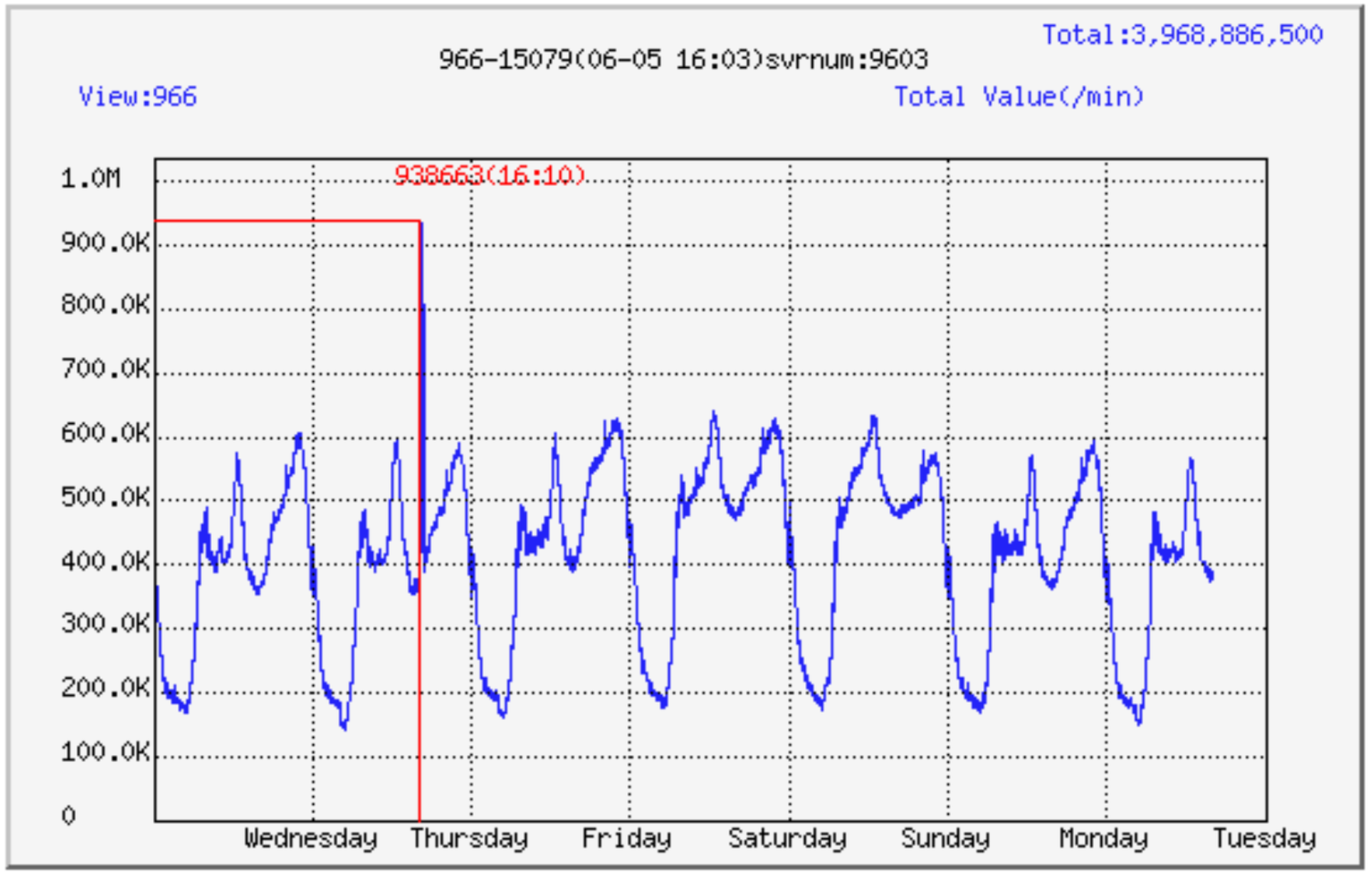
分类特征

- 熵特征
- 值分布特征
- 小波分析特征

特征工程

统计特征

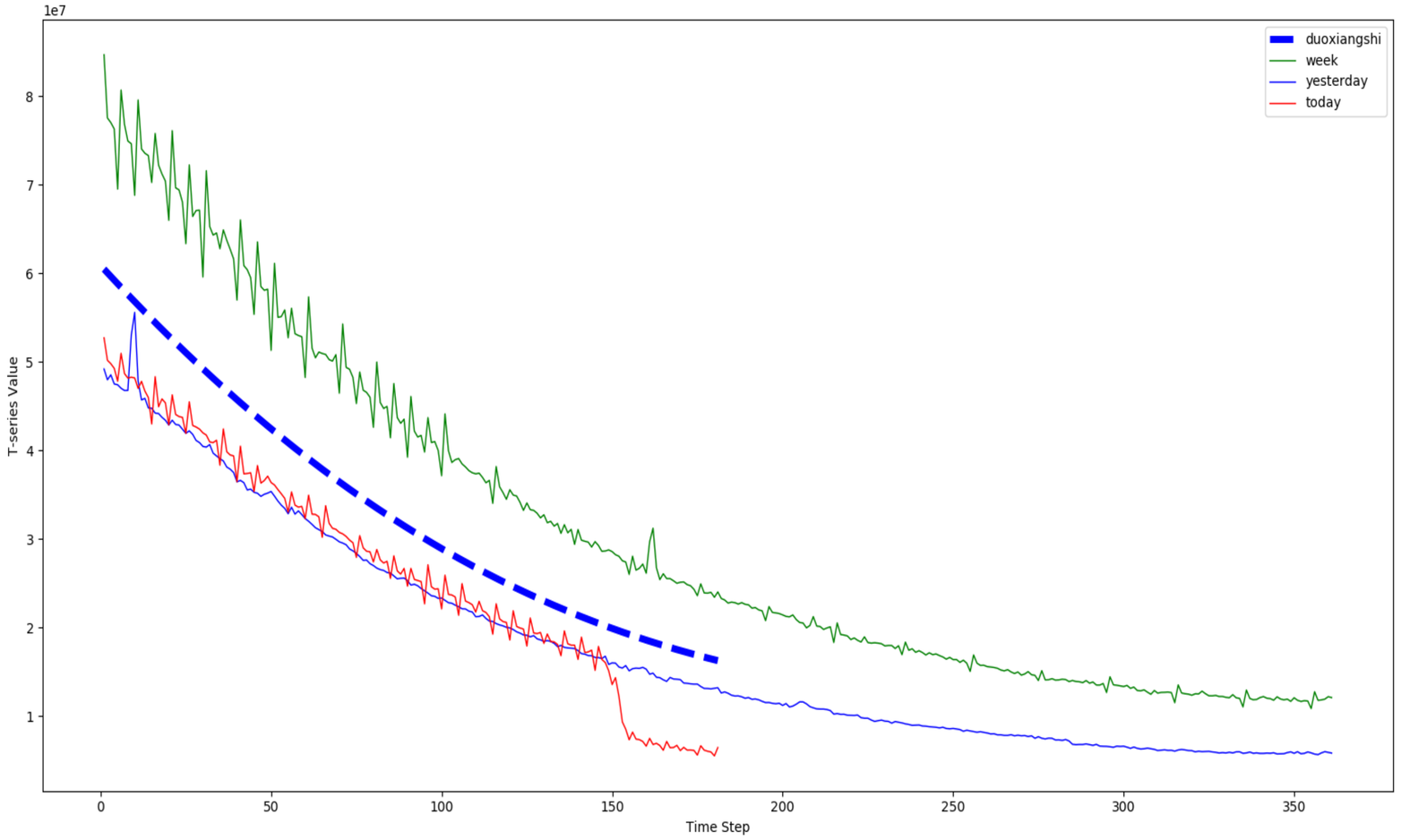
- 最大值，最小值，值域
- 最小值位置、最大值位置
- 均值，中位数
- 平方和，重复值
- 方差，偏度，峰度
- 同比，环比，周期性
- 自相关系数，变异系数



特征工程

拟合特征

- 移动平均算法
- 带权重的移动平均算法
- 指数移动平均算法
- 二次指数移动平均算法
- 三次指数移动平均算法
- 奇异值分解算法
- 自回归算法
- 深度学习算法

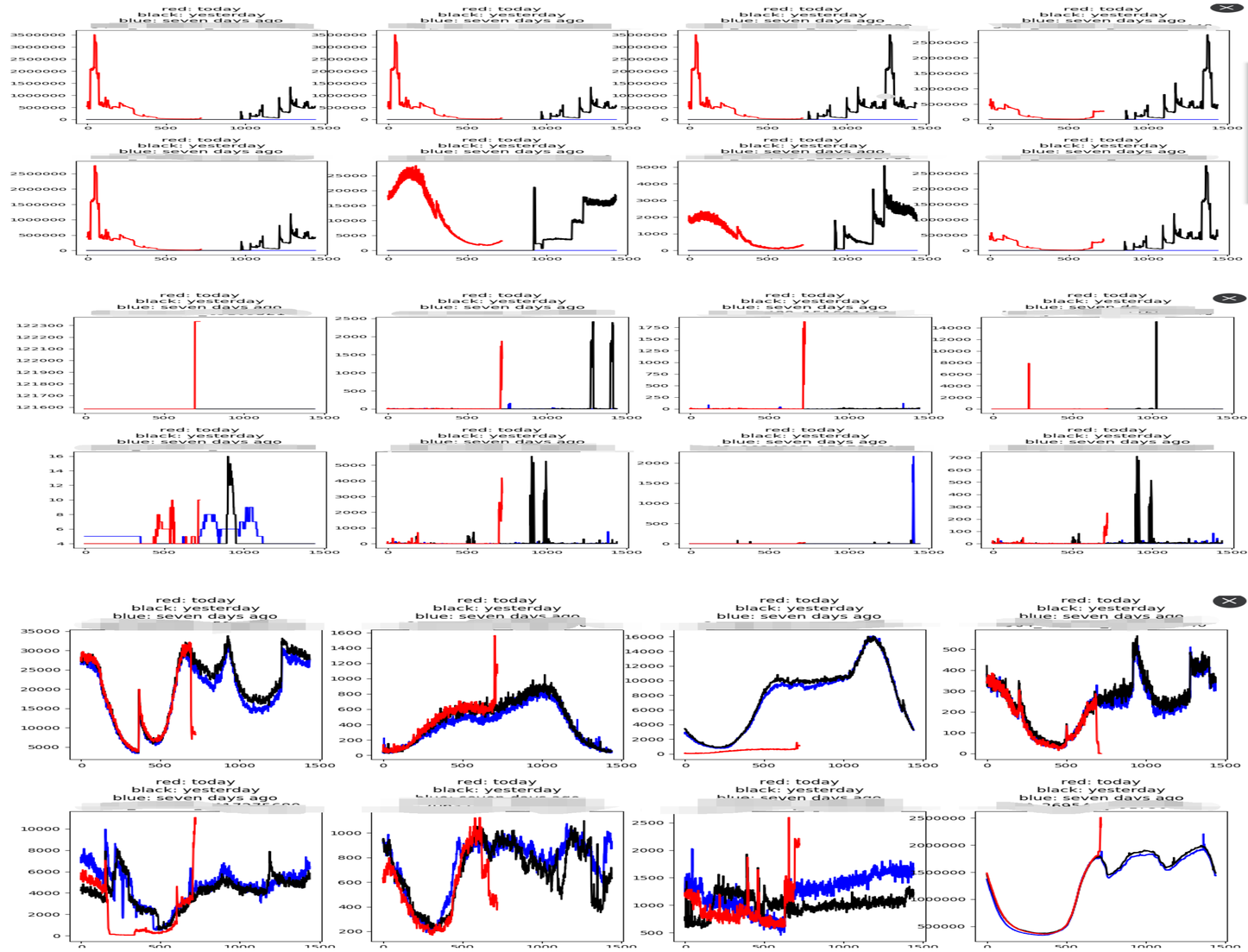


特征工程

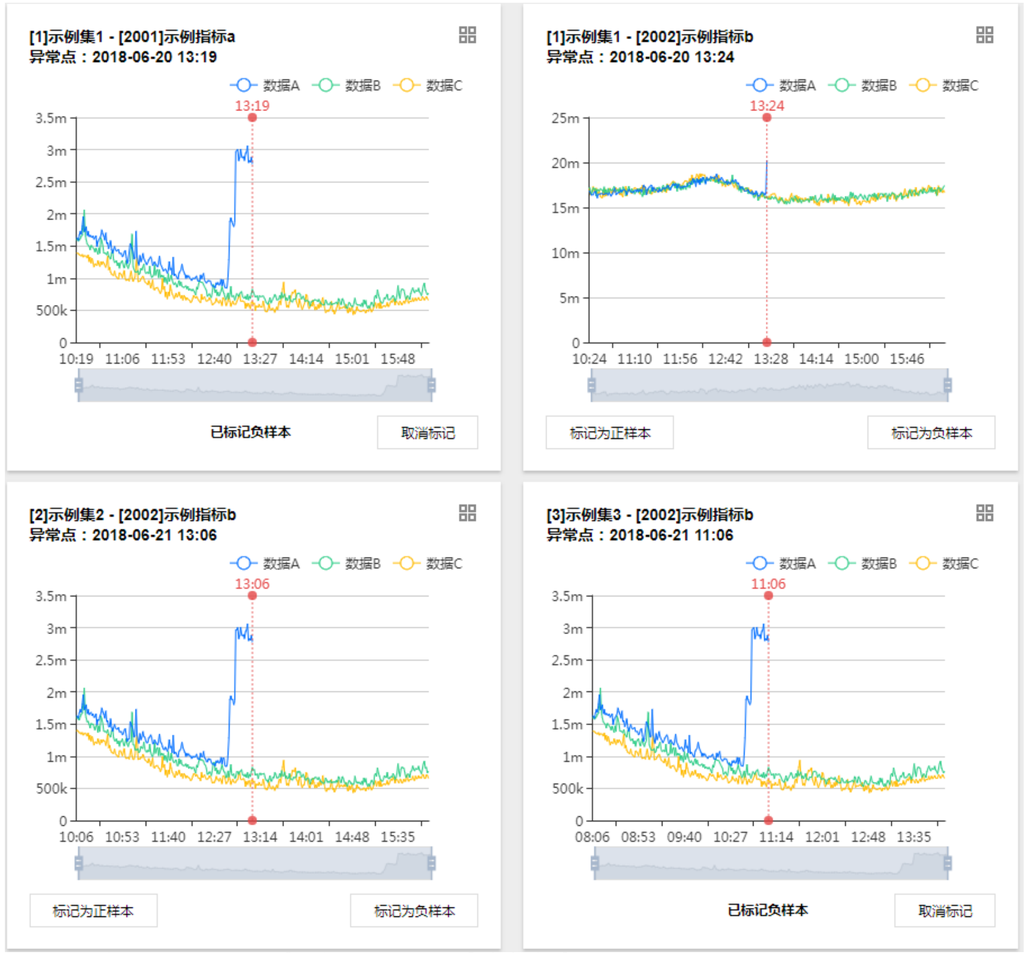
分类特征

- 熵特征
- 值分布特征
- 小波分析特征

聚类
Kmeans
分类器



打标工程



view_id	attr_id	beg_time	end_time	sender	alarm_continue	is_check
7335	2151648	2017-11-07 19:02:03	2017-11-07 19:10:02	0	0	1
1405	142637	2017-11-07 19:59:02	2017-11-07 20:06:02	0	0	1
15669	222132	2017-11-08 00:00:02	2017-11-08 10:39:02	0	0	1
6327	390161	2017-11-08 00:04:02	2017-11-08 00:11:02	0	0	1
6368	390161	2017-11-08 00:04:02	2017-11-08 00:12:03	0	0	1
6327	390161	2017-11-08 00:04:03	2017-11-08 00:12:02	0	0	1
6327	390163	2017-11-08 00:04:02	2017-11-08 00:11:02	0	0	1
6327	390163	2017-11-08 00:04:03	2017-11-08 00:12:02	0	0	1
6241	297212	2017-11-08 09:48:02	2017-11-08 09:53:02	0	0	1
6368	390163	2017-11-08 00:04:02	2017-11-08 00:12:03	0	0	1
16397	2389117	2017-11-08 09:03:03	2017-11-08 09:08:02	0	0	1
4392	20	2017-11-08 10:04:02	2017-11-08 10:09:02	0	0	1
4651	54039	2017-11-08 00:04:03	2017-11-08 00:09:03	0	0	1
11395	129854	2017-11-08 08:53:02	2017-11-08 08:58:02	0	0	1
15013	2389117	2017-11-08 09:04:03	2017-11-08 09:09:02	0	0	1
18018	629583	2017-11-08 09:20:02	2017-11-08 09:25:02	0	0	1
69141	10788	2017-10-25 11:04:03	2017-10-25 11:05:03	1	0	0
69141	10788	2017-10-25 11:04:03	2017-10-25 11:05:03	1	0	0
69141	10788	2017-10-25 11:04:03	2017-10-25 11:05:03	1	0	0
69141	10788	2017-10-25 11:04:03	2017-10-25 11:05:03	1	0	0
69141	10788	2017-10-25 11:04:03	2017-10-25 11:05:03	1	0	0
69141	10788	2017-10-25 11:04:03	2017-10-25 11:05:03	0	0	0
69141	10788	2017-10-25 11:04:03	2017-10-25 11:05:03	0	0	0

1.输出异常视图到前端页面
2.人工确认是否真的异常，假异常则校正
3.后台根据人工校正的结果，存下校正后的所有结果：正常记为1，异常记为0

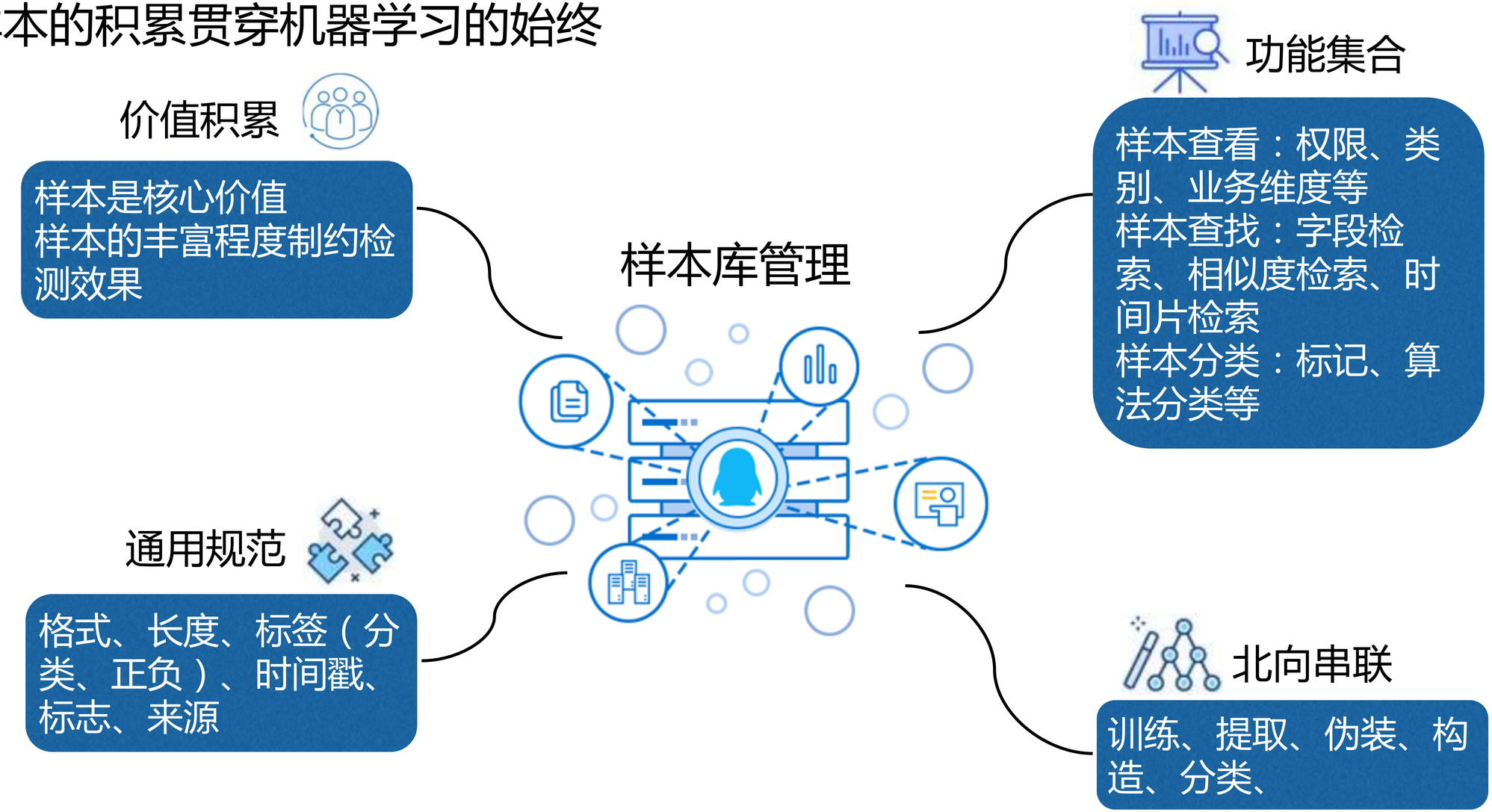
TABLE OF

CONTENTS 大纲

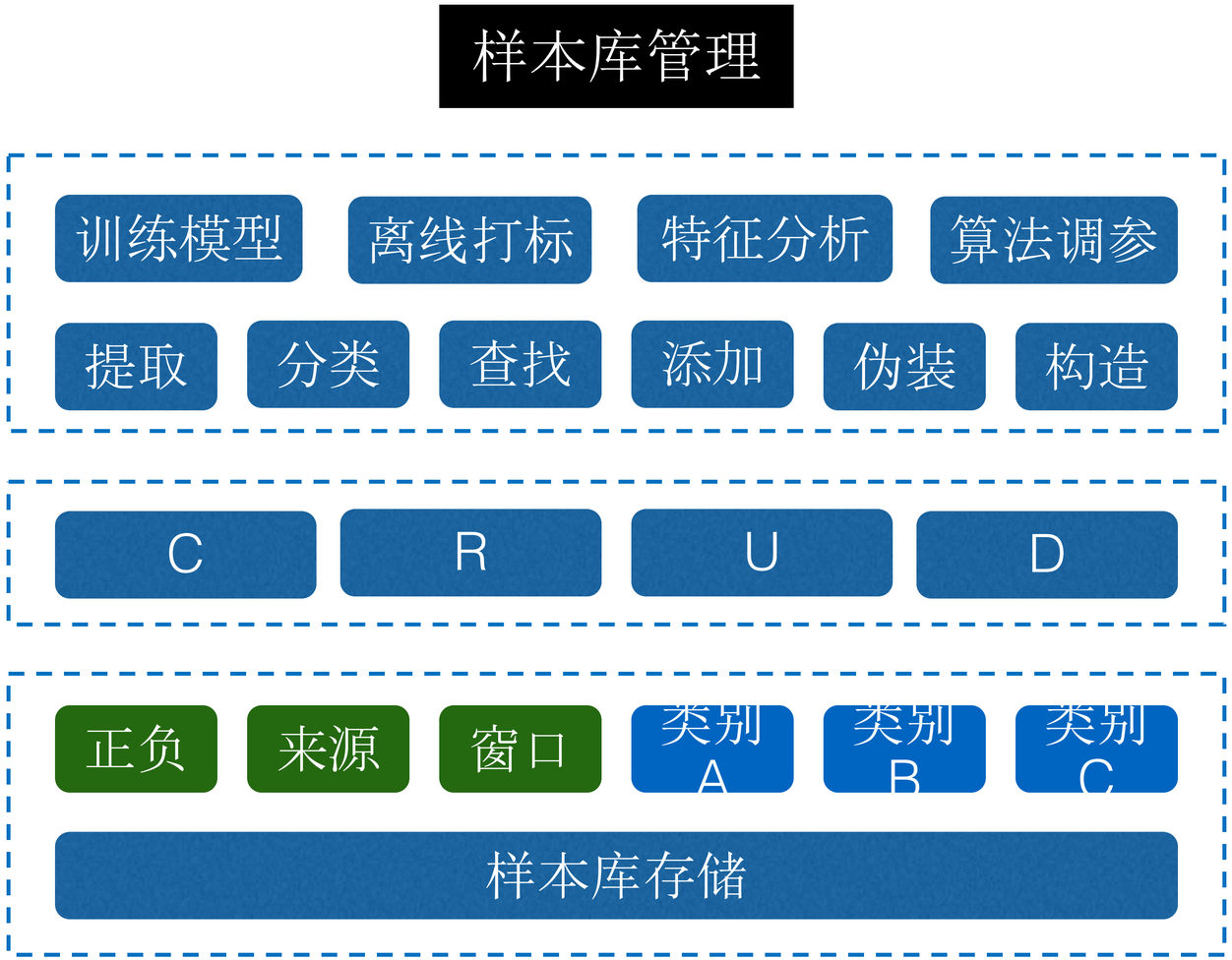
- 传统时序监控的问题与新思路
- 检测算法原理与应用
- 特征工程与打标工程
- 样本库建设与管理
- Metis概述（智能运维应用实践）

样本库管理与建设

- 样本的积累贯穿机器学习的始终



样本库管理与建设



显著提升应用效率和数据规范

Action层：触发功能与样本数据的交互

Service层：功能模块的逻辑应用实现

DAO层：封装与数据进行联络的任务，无业务逻辑

数据层：根据样本量选择存储；三级分类

Metis时间序列异常检测业务效果

目前效果

用少量模型覆盖所有曲线，统计判别+无监督+有监督

90%+

准确率

计算方法：人工抽查，查看告警出来的时间序列和时间点是否准确

辅助工具：打标工程

80%+

召回率

计算方法：人工从业务中选择一批异常的时间序列和相应的时间点，然后让这批序列通过现有模型，看看是否被召回

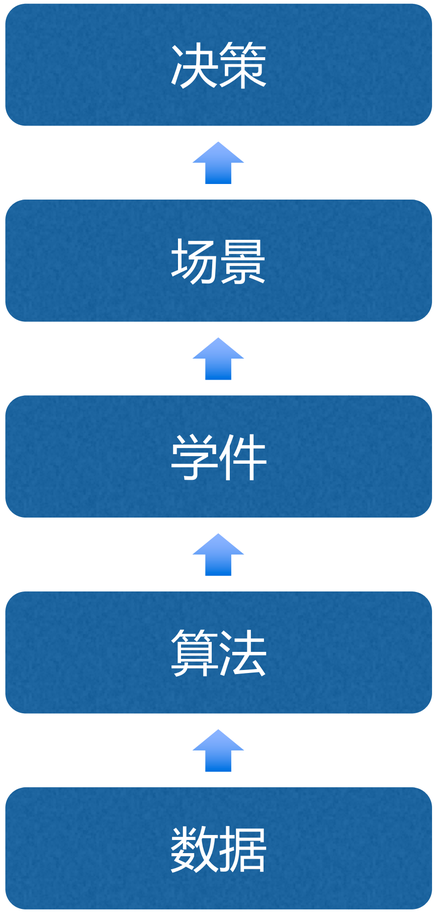
辅助工具：样本库管理

TABLE OF

CONTENTS 大纲

- 传统时序监控的问题与新思路
- 检测算法原理与应用
- 特征工程与打标工程
- 样本库建设与管理
- Metis概述（智能运维应用实践）

渐进式的AIOps能力



分层 明细

多目标自调整AIOPS大脑

主场景均实现流程化免干预AI运维能力

串联AI运维应用如故障自愈

效率	质量	成本
AI运维组件(学件)库 智能预测 智能变更 问答机器人 变更机器人 建议机器人 ...	AI运维组件(学件)库 异常检测 异常预测 根因分析 舆情监控 ...	AI运维组件(学件)库 性能调优 成本分析 强化改进 碎片整理 容量管理 ...

运维开发框架及工具库

AI算法平台和算法库

离线计算 | 实时计算

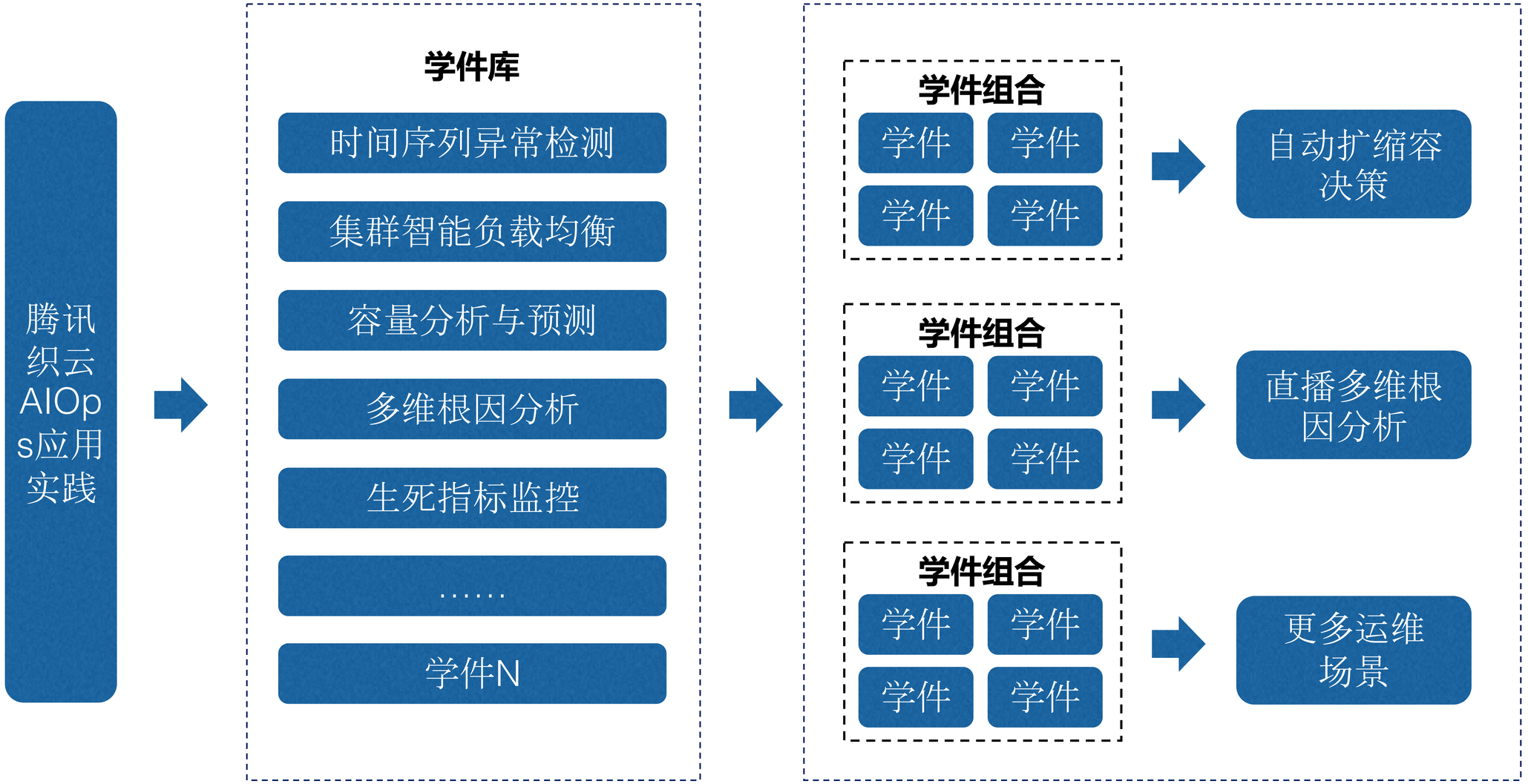
元数据管理、数据仓库

数据清洗、ETL、特征工程

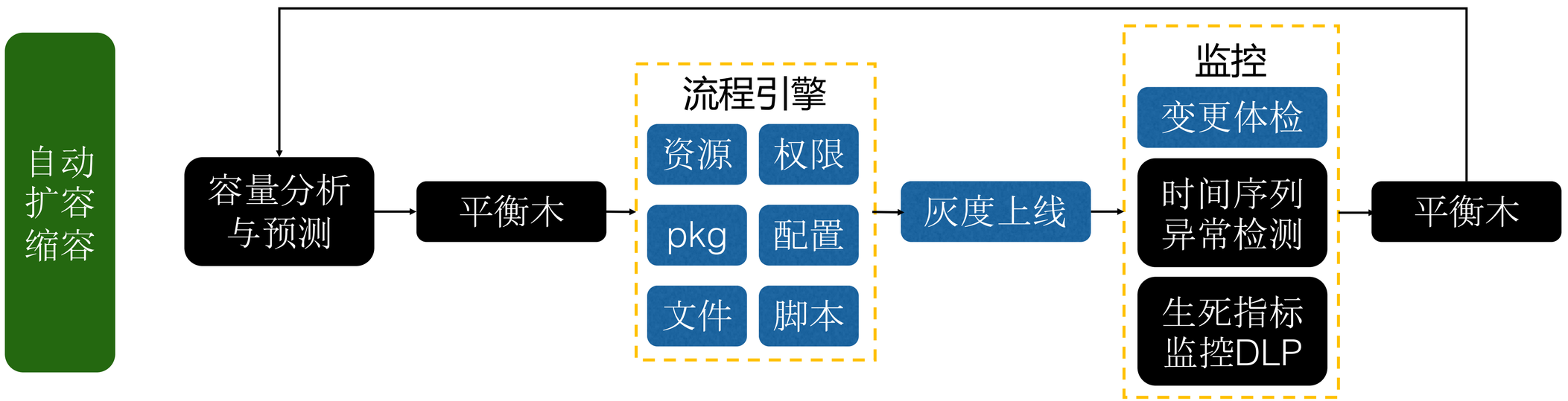
数据通道

数据采集如各类agent

织云Metis


腾讯织云AIOps应用实践
学件库
时间序列异常检测
集群智能负载均衡
容量分析与预测
多维根因分析
生死指标监控
……
学件N
学件组合
学件
学件
学件
学件
自动扩缩容决策
学件组合
学件
学件
学件
学件
直播多维根因分析
学件组合
学件
学件
学件
学件
更多运维场景

串联应用案例


自动
扩容
缩容
容量分析
与预测
平衡木
流程引擎
资源
权限
pkg
配置
文件
脚本
灰度上线
监控
变更体检
时间序列
异常检测
生死指标
监控DLP
平衡木


绿色表示运维场景
蓝色表示自动化工具
黑色表示智能化学件


流程管理 / 流程执行
获取参数列表
检查一致性上报
检查设备连通性
获取资源配置
申请权限
屏蔽告警
L5主调扩容
cmlb主调扩容
安装程序包
同步文件
获取cc参数
发配置
执行脚本
启动软件包
进程端口扫描
告警屏蔽解除
执行测试工具
L5被调扩容
CMLB被调扩容
上报变更日志

ArchSummit
全 球 架 构 师 峰 会

关注 ArchSummit 公众号

获取国内外一线架构设计

了解上千名知名架构师的实践动向

北京站：2018年12月7–10日

极客时间

极客时间企业账户

一种全新的团队学习方式

碎片时间，提升研发团队整体战斗力

扫码解锁更多


Metis全新发布
织云AIOps体系，智能运维应用实践