AIOps Challenge

微服务架构电商系统下故障识别和分类

中南-天云战队/杨平

中南大学、天云软件技术有限公司

2022 CCF国际AIOps挑战赛决赛暨AIOps研讨会

AIOps Challenge

2022 CCF国际AIOps挑战赛决赛
暨AIOps研讨会

目录
CONTENTS

Tcloudsoft

1 团队介绍

AIOps Challenge

中南-天云战队 =  + Tcloudsoft®

参赛人员：

- 杨平（博士研究生、CTO）
- 刘鑫霖（算法经理）
- 邓涵宇（硕士研究生）
- 李君健（硕士研究生）
- 陈梓豪（硕士研究生）
- 赵琳（算法总监）
- 黄载群（算法工程师）
- 蒋颖（数据经理）

参赛单位：

- 中南大学
- 天云软件技术有限公司

指导老师：

- 王建新（教授、博导）
- 黄家玮（教授、博导）

比赛成绩：

- 复赛第四名

Tcloudsoft®

2 数据分析-数据概览

AIOps Challenge

本次竞赛一共有3大类数据：指标、日志和调用链，其主要来自于微服务系统所收集的多模态监控数据。
其中，指标数据又包括业务指标和性能指标，业务指标共有4个，性能指标近400个。

为保证算法模型的通用性和可操作性，本比赛最终方案主要采用了**性能指标数据**

业务指标

时间戳	系统响应率	业务成功率	交易量	平均响应时间	服务名称
timestamp	rr	sr	count	mrt	service
1611224141	1	0.99	8000	800	adservice

性能指标

时间戳	指标所在对象	指标名	指标值
timestamp	cmdb_id	kpi_name	value
1611224141	node-1	system.cpu.iowait	80

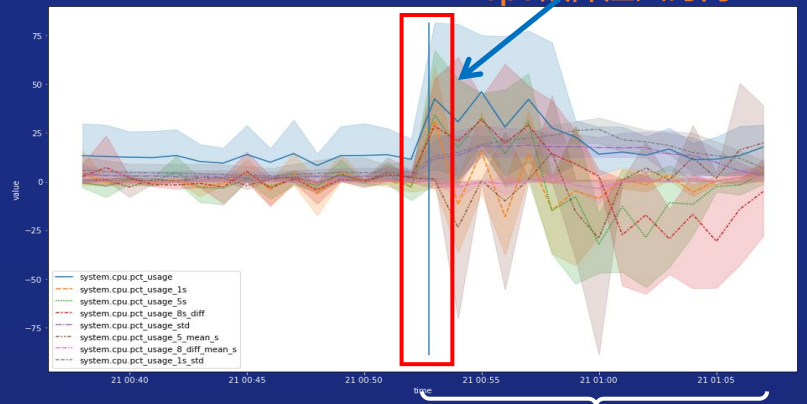
Tcloudsoft

2 数据分析-故障相关指标影响

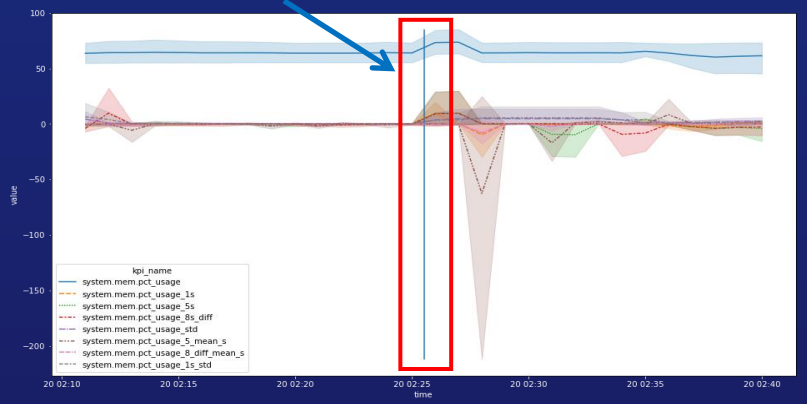
AIOps Challenge

node级故障数据呈现

CPU故障

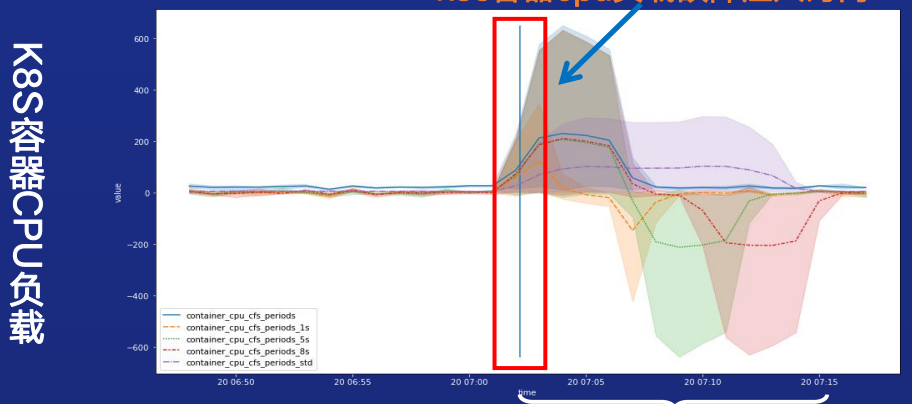


内存消耗

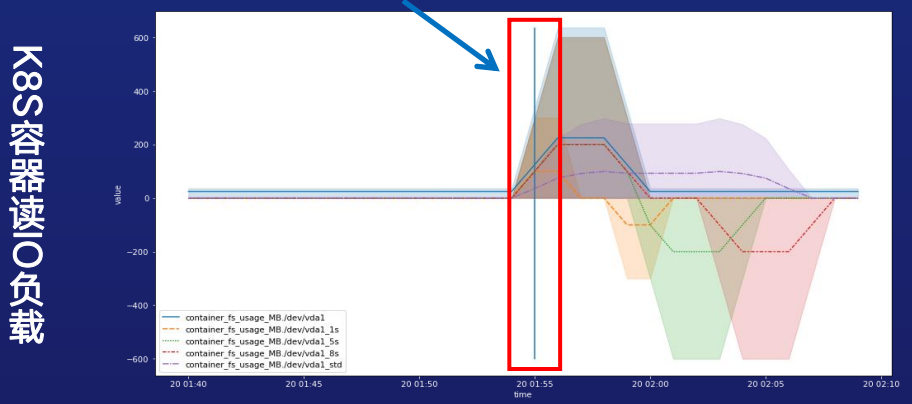


pod级故障数据呈现

K8S容器CPU负载



K8S容器读IO负载

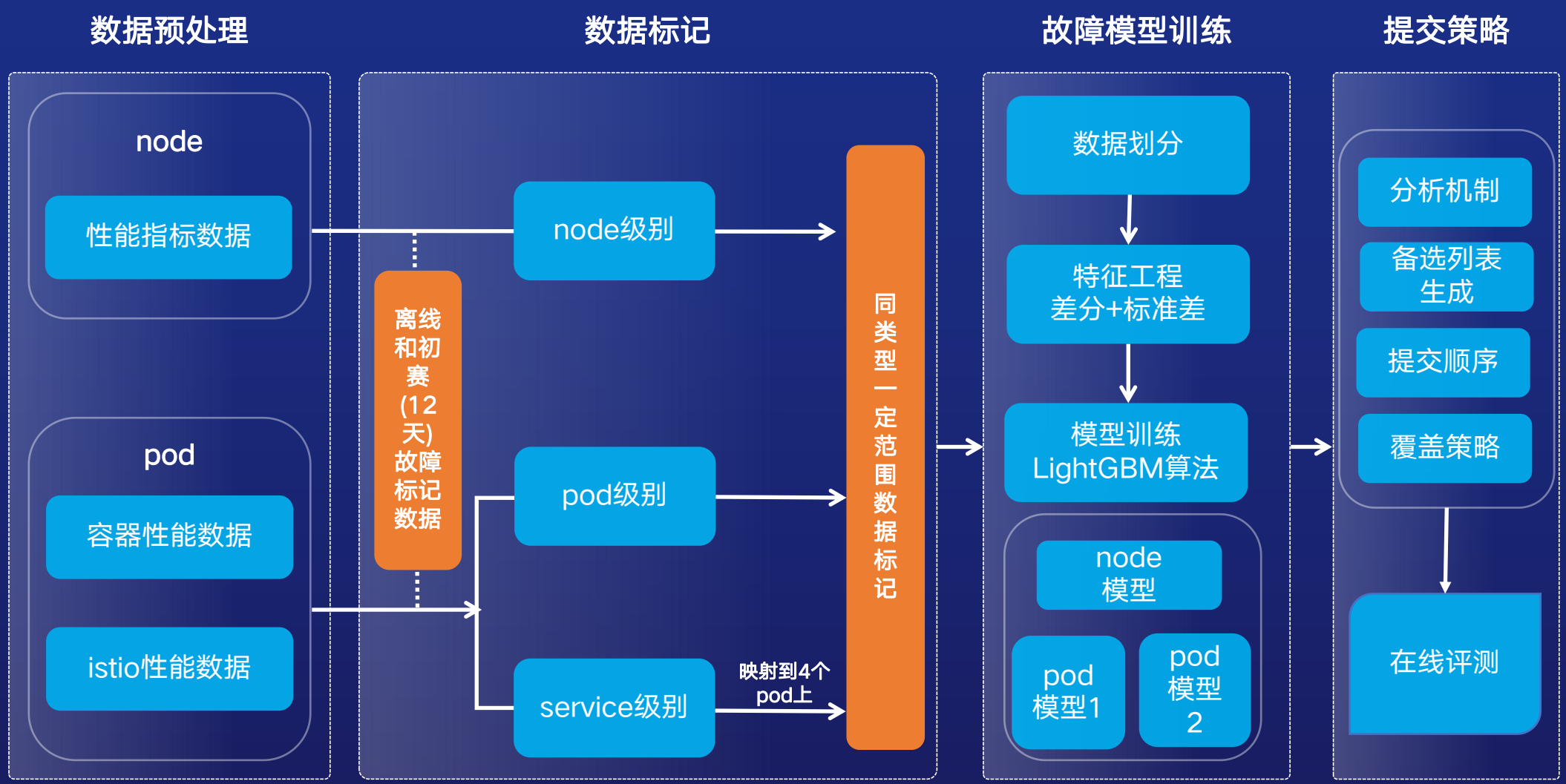


Tcloudsoft

3 方案介绍-方案框架

AIOps Challenge

本方案主要分为数据预处理，数据标记，故障模型训练，提交策略四个模块：



Tcloudsoft

3 方案介绍-数据预处理

AIOps Challenge

对于node性能指标数据，将原始数据对node-1至node-6按照1分钟级别进行合并

Node性能数据

timestamp	cmdb_id	kpi_name	value
1647619200	node-2	system.io.avg_q_sz	0.16
1647619200	node-5	system.io.avg_q_sz	0.01
1647619200	node-1	system.io.avg_q_sz	0
1647619200	node-6	system.io.avg_q_sz	0
1647619200	node-3	system.io.avg_q_sz	0
1647619200	node-4	system.io.avg_q_sz	0
1647619260	node-3	system.io.avg_q_sz	0
1647619260	node-5	system.io.avg_q_sz	0.01
1647619260	node-4	system.io.avg_q_sz	0
1647619260	node-6	system.io.avg_q_sz	0.01

timestamp	cmdb_id	kpi_name	value
1647619200	node-2	system.mem.real.pct_useage	50.05
1647619200	node-1	system.mem.real.pct_useage	80.78
1647619200	node-5	system.mem.real.pct_useage	45.89
1647619200	node-4	system.mem.real.pct_useage	46.51
1647619200	node-6	system.mem.real.pct_useage	19.7
1647619200	node-3	system.mem.real.pct_useage	50.6
1647619260	node-1	system.mem.real.pct_useage	80.69
1647619260	node-6	system.mem.real.pct_useage	20.5
1647619260	node-5	system.mem.real.pct_useage	45.41
1647619260	node-3	system.mem.real.pct_useage	50.59

timestamp	cmdb_id	kpi_name	value
1647619200	node-4	system.net.udp.in_datagrams	507.2
1647619200	node-5	system.net.udp.in_datagrams	608.1
1647619200	node-6	system.net.udp.in_datagrams	1604.83
1647619200	node-1	system.net.udp.in_datagrams	400.85
1647619200	node-2	system.net.udp.in_datagrams	947.17
1647619200	node-3	system.net.udp.in_datagrams	2236.55
1647619260	node-6	system.net.udp.in_datagrams	1457.48
1647619260	node-5	system.net.udp.in_datagrams	695.82
1647619260	node-1	system.net.udp.in_datagrams	547.97
1647619260	node-2	system.net.udp.in_datagrams	1183.18

.....

timestamp	cmdb_id	kpi_name	value
1647705600	node-5	system.disk.free	4088185173
1647705600	node-1	system.disk.free	3243289856
1647705600	node-6	system.disk.free	2591089920
1647705600	node-3	system.disk.free	3271603712
1647705600	node-2	system.disk.free	3274320384
1647705600	node-4	system.disk.free	3264192256
1647705660	node-2	system.disk.free	3273836288
1647705660	node-5	system.disk.free	4087871829
1647705660	node-6	system.disk.free	2590732544
1647705660	node-1	system.disk.free	3242775808

timestamp	cmdb_id	kpi_name	value
1647705600	node-6	ping.can_connect	1
1647705600	node-2	ping.can_connect	1
1647705600	node-4	ping.can_connect	1
1647705660	node-4	ping.can_connect	1
1647705660	node-2	ping.can_connect	1
1647705660	node-6	ping.can_connect	1
1647705720	node-4	ping.can_connect	1
1647705720	node-6	ping.can_connect	1
1647705720	node-2	ping.can_connect	1

node级别

timestamp	cmdb_id	ping.can_connect	system.cpu.iowait	system.cpu.pct_usage	system.cpu.system	system.cpu.user	system.disk.free	system.disk.pct_usage	system.disk.readonly	...	system.proces
1652025600	node-1	1.0	0.71	28.87	2.20	25.97	3.098895e+09	46.33	0.0	...	0.0
1652025600	node-2	1.0	1.69	3.60	0.96	0.96	3.130082e+09	41.07	0.0	...	0.0
1652025600	node-3	NaN	1.11	6.80	1.78	3.91	3.127453e+09	41.07	0.0	...	0.0
1652025600	node-4	1.0	1.71	3.62	1.12	0.79	NaN	NaN	NaN	...	0.0
1652025600	node-5	NaN	0.00	7.27	2.02	5.25	3.895508e+09	68.26	0.0	...	11.0
...	...	...	...	...	...	...	...	...	...	...	...
1652111940	node-2	1.0	4.53	9.09	1.24	3.32	3.057707e+09	41.36	0.0	...	0.0
1652111940	node-3	NaN	0.50	3.82	1.38	1.94	3.055192e+09	41.78	0.0	...	0.0
1652111940	node-4	1.0	0.00	5.49	1.95	1.16	NaN	NaN	NaN	...	0.0
1652111940	node-5	NaN	0.00	4.82	2.08	2.75	3.797746e+09	68.83	0.0	...	17.0
1652111940	node-6	1.0	0.02	17.46	10.07	7.37	1.983716e+09	26.67	0.0	...	197.0

性能指标字段

合并后node对应的cmdb_id

容器性能数据

istio性能数据

pod级别

timestamp	cmdb_id	container_cpu_cfs_periods	container_cpu_system_seconds	container_cpu_usage_seconds	container_cpu_user_seconds	container_file_descriptors	container_fs_limit
1647705600	adservice-2	117.000000	0.020000	0.068313	0.035000	93.0	604630.738281
1647705600	cartservice2-0	101.666667	0.023333	0.069902	0.033333	162.0	604630.738281
1647705600	checkoutservice-2	1.500000	0.005000	0.015124	0.010000	10.0	604630.738281
1647705600	frontend-1	90.333333	0.076667	0.257937	0.110000	72.0	604630.738281
1647705600	frontend-2	93.000000	0.070000	0.231060	0.120000	19.0	604630.738281
...	...	...	...	...	...	...	...
1647791940	recommendationservice-1	167.000000	0.025000	0.149284	0.095000	11.0	604630.738281
1647791940	recommendationservice2-0	168.000000	0.020000	0.102714	0.080000	10.0	604630.738281
1647791940	shippingservice-0	13.500000	0.000000	0.007780	0.005000	9.0	604630.738281
1647791940	shippingservice-1	12.000000	0.000000	0.012291	0.010000	10.0	604630.738281
1647791940	shippingservice2-0	27.000000	0.010000	0.019877	0.010000	9.0	604630.738281

性能指标字段

合并后pod对应的cmdb_id

cloudsoft

3 方案介绍-node和pod故障标记

AIOps Challenge

node故障标记

总体原则：结合给定的12天故障标记数据，将相同cmdb_id故障注入后8分钟内的数据标记为相同故障。

故障类别映射

故障类别	标签
normal	0
node 内存消耗	1
node 磁盘读IO消耗	2
node 磁盘写IO消耗	3
node节点CPU故障	4
node节点CPU爬升	5
node磁盘空间消耗	6

举例说明

故障数据

timestamp	level	cmdb_id	failure_type
1647743133	node	node-1	node 内存消耗

待标记数据

timestamp	cmdb_id	……	failure_type
1647743160	node-3	……	0
1647743160	node-1	……	1
……	……	……	……
1647743700	node-1	……	0

pod|service故障标记

总体原则：service级别是将其映射到4个pod上，再依据pod标记方案对其进行标记。

故障类别映射

故障类别	标签
normal	0
k8s容器网络资源包损坏	1
k8s容器cpu负载	2
k8s容器网络丢包	3
k8s容器内存负载	4
k8s容器网络延迟	5
k8s容器进程中止	6
k8s容器读io负载	7
k8s容器网络资源包重复发送	8
k8s容器写io负载	9

举例说明

故障数据

timestamp	level	cmdb_id	failure_type
1647749457	service	recommendationservice	k8s容器cpu负载

待标记数据

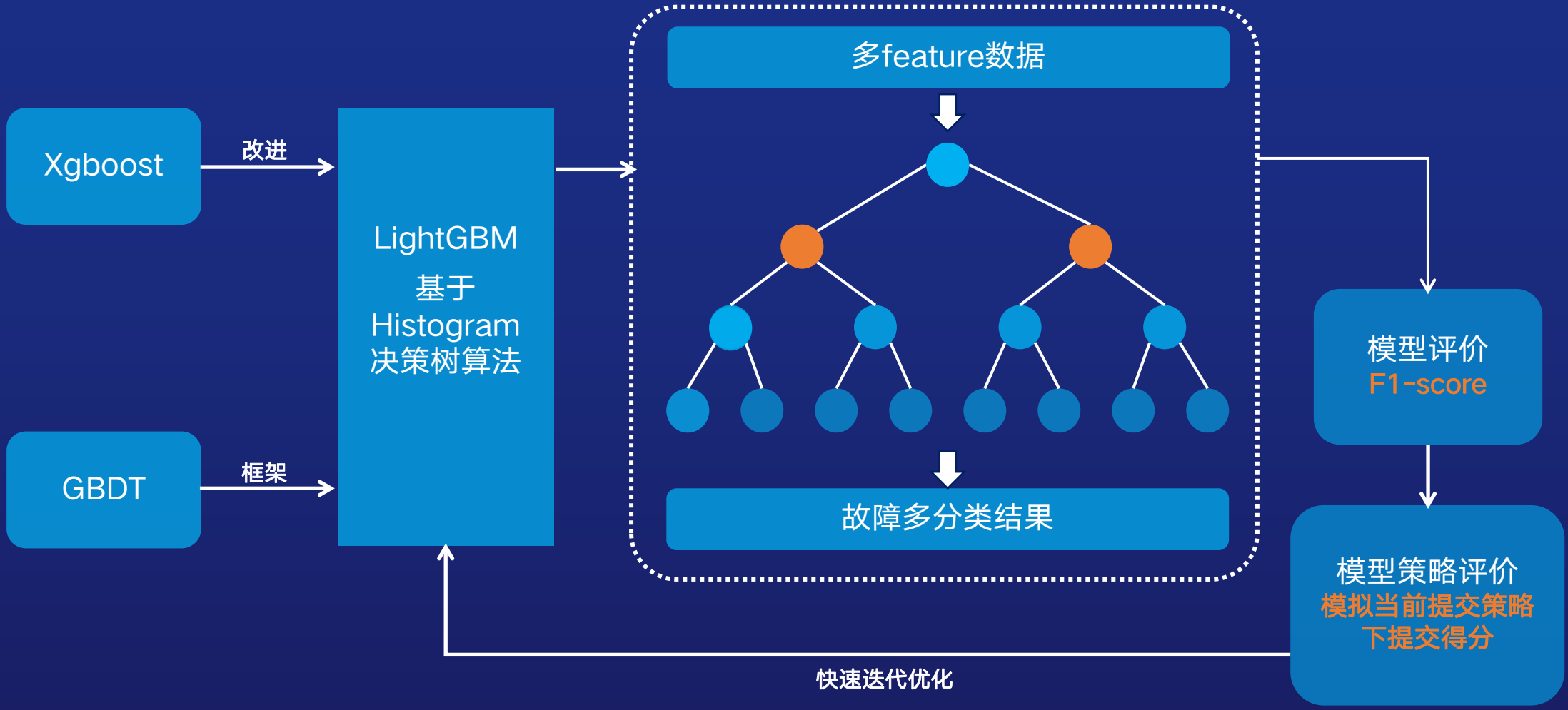
timestamp	cmdb_id	……	failure_type
1647749460	recommendationservice-0	……	2
1647749460	shippingservice-1	……	0
……	……	……	……
1647750020	recommendationservice-0	……	0

Tcloudsoft

3 方案介绍-故障识别模型构建算法

AIOps Challenge

故障识别模型构建算法：使用LightGBM算法构建故障识别模型（即多分类模型）



Tcloudsoft

3 方案介绍-node故障识别模型

AIOps Challenge

特征衍生

- ✓ 一阶差分
- ✓ 5分钟滑动窗口求和的差分
- ✓ 8分钟滑动窗口求和的差分
- ✓ 8分钟滑动窗口的标准差
- ✓ 8分钟滑动窗口均值的增长率
- ✓ 8分钟滑动窗口求和的增长率
- ✓ 一阶差分8分钟滑动窗口的标准差

```
df[[col + '_1s' for col in pod_data_col_list]] = df[pod_data_col_list].diff()
df[[col + '_5s' for col in pod_data_col_list]] = df[pod_data_col_list].rolling(5, min_periods=1).sum().diff()
df[[col + '_8s' for col in pod_data_col_list]] = df[pod_data_col_list].rolling(8, min_periods=1).sum().diff()
df[[col + '_std' for col in pod_data_col_list]] = df[pod_data_col_list].rolling(8, min_periods=1).std()
bx =  df[pod_data_col_list].rolling(8, min_periods=1).mean().diff()
bx_sum =  df[pod_data_col_list].rolling(8, min_periods=1).sum().diff()
df[[col + '_8_mean_s' for col in pod_data_col_list]] = bx / bx.shift(-1)
df[[col + '_8_sum_s' for col in pod_data_col_list]]= bx_sum / bx_sum.shift(-1)
df[[col + '_1s_std' for col in node_col]] = df[[col + '_1s' for col in node_col]].rolling(8, min_periods=1).std()
```

衍生去除无用特征

特征筛选

- ✓ 衍生原始特征数：44，衍生特征加原始特征数：365
- ✓ 训练集数据量：60480 ,其中故障标签数据：680
- ✓ 测试集数据量：43200，其中故障标签数据：536

```
'ping.can_connect',
'system.mem.free',
'system.swap.used',
'system.mem.total',
'system.swap.free',
'system.net.udp.snd_buf_errors',
'system.swap.si',
'system.net.packets_in.error',
'system.swap.total',
'system.os.nofile.max',
'system.swap.used_pct',
'system.net.udp.rcv_buf_errors',
'system.swap.so',
'system.net.packets_out.error',
'system.disk.readonly'
```

模型训练

模型训练

使用LightGBM算法，经调优后参数设定如下：

bagging_freq: 1	objective: multiclass	min_sum_hessian_in_leaf: 6
num_class: 7	boosting: gbdt	bagging_fraction: 0.7
num_leaves: 6	min_data_in_leaf: 30	bagging_seed: 111
max_depth: 6	learning_rate: 0.1	random_state: 13
lambda_l1: 0.25	feature_fraction: 0.8	lambda_l2: 0.5

模型效果评估

模型评估

模型在数据划分第二阶段中测试集上的评价表现如下：

七分类模型在测试集中的平均F1分数为0.76

	precision	recall	f1-score	support
0	1.00	1.00	1.00	42664
1	0.97	0.40	0.57	72
2	0.75	0.69	0.72	96
3	0.84	0.73	0.78	88
4	0.90	0.82	0.86	112
5	0.74	0.36	0.48	56
6	0.94	0.92	0.93	112
accuracy			1.00	43200
macro avg	0.88	0.70	0.76	43200
weighted avg	0.99	1.00	0.99	43200

宏平均F1分数值

Tcloudsoft

3 方案介绍-pod故障识别模型

AIOps Challenge

特征衍生

- 一阶差分
- 5分钟滑动窗口求和的差分
- 8分钟滑动窗口求和的差分
- 8分钟滑动窗口的标准差

```
df[[col + '_1s' for col in pod_data_col_list]] = df[pod_data_col_list].diff()
df[[col + '_5s' for col in pod_data_col_list]] = df[pod_data_col_list].rolling(5, min_periods=1).sum().diff()
df[[col + '_8s' for col in pod_data_col_list]] = df[pod_data_col_list].rolling(8, min_periods=1).sum().diff()
df[[col + '_std' for col in pod_data_col_list]] = df[pod_data_col_list].rolling(8, min_periods=1).std()
```

特征筛选

衍生去除部分噪音特征

- 衍生原始特征数：162，衍生后特征加原始特征数：810
- 训练集数据量：403192，其中故障标签数据：4161
- 测试集数据量：302317，其中故障标签数据：3485

```
'container_last_seen',
'istio_request_duration_milliseconds.grpc.200.0.0',
'istio_request_messages',
'istio_requests.grpc.200.0.0',
'istio_response_bytes.grpc.200.0.0',
'istio_response_messages',
'container_cpu_load_average_10s',
'container_fs_inodes_free./dev/vda1',
'container_fs_io_current./dev/vda1',
'container_fs_io_time_seconds./dev/vda1',
'container_fs_io_time_weighted_seconds./dev/vda1',
'container_fs_limit_MB./dev/vda1',
'container_fs_read_seconds./dev/vda1',
'container_fs_reads./dev/vda1',
'container_fs_reads_merged./dev/vda1',
'container_fs_sector_reads./dev/vda1',
'container_fs_sector_writes./dev/vda1',
'container_fs_write_seconds./dev/vda1',
'container_fs_writes./dev/vda1',
'container_fs_writes_merged./dev/vda1',
'container_memory_swap',
'container_network_receive_errors.eth0',
'container_network_transmit_errors.eth0',
'container_network_transmit_packets_dropped.eth0',
'container_spec_cpu_period',
'container_spec_memory_reservation_limit_MB',
'container_tasks_state.iowaiting',
'container_tasks_state.running',
'container_tasks_state.sleeping',
'container_tasks_state.stopped',
'container_tasks_state.uninterruptible',
'container_threads_max'
```

标记处为噪音指标，其余为0值或单一值指标

Tcloudsoft

3 方案介绍-pod故障识别模型

AIOps Challenge

模型训练

使用LightGBM算法，经调优后参数设定如下：

bagging_freq: 1	objective: multiclass
num_class: 10	boosting: gbdt
num_leaves: 6	min_data_in_leaf: 30
max_depth: 6	learning_rate: 0.1
lambda_l1: 0.25	lambda_l2: 0.5

min_sum_hessian_in_leaf: 6

bagging_fraction: 0.5

bagging_seed: 11

random_state: 2022

feature_fraction: 0.8

训练结果

模型评估

模型 1：

采用全量container加部分istio性能数据进行模型训练，其特征数共计251个。在数据划分第二阶段中，该十分类模型在测试集上的评价表现如右图所示，其F1分数为0.63

	precision	recall	f1-score	support
0	0.99	1.00	1.00	297864
1	0.57	0.27	0.36	960
2	0.98	0.67	0.80	540
3	0.45	0.50	0.48	450
4	0.89	0.62	0.73	525
5	0.70	0.58	0.64	467
6	0.69	0.46	0.55	323
7	0.98	0.68	0.80	576
8	0.15	0.35	0.21	235
9	0.91	0.61	0.73	377
accuracy			0.99	302317
macro avg	0.73	0.57	0.63	302317
weighted avg	0.99	0.99	0.99	302317

模型 2：

采用全量container性能数据进行模型训练，其特征数共计198个。在数据划分第二阶段中，该十分类模型在测试集上的评价表现如右图所示，其F1分数为0.59

	precision	recall	f1-score	support
0	0.99	1.00	1.00	298832
1	0.81	0.29	0.43	448
2	0.76	0.91	0.83	372
3	0.68	0.20	0.31	496
4	0.75	0.67	0.71	368
5	0.62	0.04	0.07	392
6	0.62	0.48	0.54	216
7	0.85	0.93	0.89	399
8	0.69	0.14	0.23	540
9	0.94	0.80	0.86	254
accuracy			0.99	302317
macro avg	0.77	0.55	0.59	302317
weighted avg	0.99	0.99	0.99	302317

pod故障识别双模型机制：模型1检测网络丢包和网络延迟2类故障，模型2检测其它8类故障

Tcloudsoft

3 方案介绍-提交策略

AIOps Challenge

模型分类结果

timestamp	cmdb_id	pre_y
1647743160	node-3	0
1647743160	node-1	1
.....	.....	.....
1647749460	frontend-2	0

对node、pod、service分别进行分析

分析机制

- ✓ 10秒钟进行一次检测
- ✓ 提取当前时间前10分钟的检测结果进行分析

node 级别故障

历史记录中，相同cmdb_id出现次数大于等于1次的，将该检测结果提交至node级别备选列表

pod 级别故障

历史记录中，相同cmdb_id出现次数大于等于1次的，将该检测结果提交至pod级别备选列表

service 级别故障

如果该service下的4个pod中大于等于2个出现相同故障，结果提交至service级别备选列表

限制策略

- 10分钟内检出的pod级别故障不允许覆盖node和service故障；
- 10分钟内检出的node故障不允许覆盖service故障；
- 对于同一cmdb，20分钟内不可重复提交；
- 对pod级别的k8s容器网络资源包损坏、k8s容器网络丢包、k8s容器网络资源包重复发送这三个故障在10分钟内检出2次才会加入提交备选列表；

node检出准确率高

service检出门槛高

按顺序提交 service>pod>node

cmdb_id	failure_type
adservice	k8s容器cpu负载
adservice-2	k8s容器cpu负载
node-1	node节点CPU爬升
adservice-1	k8s容器cpu负载
.....	.....

备选列表

提交

cmdb_id	failure_type
adservice	k8s容器cpu负载

提交结果

Tcloudsoft

4 改进思路

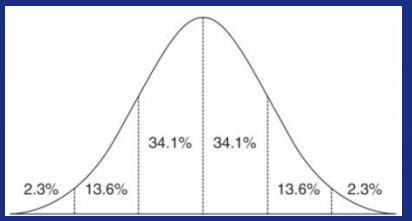
AIOps Challenge

01 尝试方案介绍：

往期正常数据 → 计算前后长短时间 → z-score

trace数据 → 计算duration均值 → z-score

node/pod模型 → Xgboost算法训练 → 评价分数没有LGB高



02 待提升的方向：

数据利用
- metric,trace,log对故障的影响

关联分析
- 建立pod和node之间的相关性

特征工程
- 面向故障针对性衍生特征

提交策略优化
- 适应检测精度调整提交顺序

Tcloudsoft

AIOps Challenge

2022 CCF国际AIOps挑战赛决赛暨AIOps研讨会

THANKS