大规模网络智能运维最佳实践

讲师简介

李忠良
中兴通讯高级系统架构师

李忠良，目前就职于中兴通讯数据智能平台系统部，中兴通讯高级系统架构师，中兴通讯网络智能化算法负责人，具有十多年一线开发经验，主要研究方向为数据库智能运维、网络智能化等方向。

背景

2025年达到自智网络L4等级成为行业共识

目录

- 数据中心能耗预测优化最佳实践
- 复杂网络故障根因分析最佳实践

案例简介

数据中心能耗预测优化：基于强化学习技术的AI节能系统实现数据中心全生命周期智慧节能，使数据中心始终运行在能效最优的状态，为客户带来直接收益。

故障根因智能分析：基于图神经网络技术实现故障根因分析，使大规模网络场景出现告警风暴时能够快速定位故障根因节点，节省运维时间降低运维成本。

数据中心能耗预测优化-案例背景

问题

2018年在中国数据中心耗电量占全社会的**2.35%**，数据中心能耗成为行业重点关注问题。

各地政府纷纷出台限耗政策，**PUE**低于**1.3**已经成为新建数据中心的入场券。

传统调优方式只能凭借运维专家经验，对**3~5**个制冷参数进行调节，难以获得最优控制策略。

$$PUE = (P_{IT} + P_{制冷} + P_{供电} + P_{其他}) / P_{IT}$$

方案

使用**AI**技术不断更新迭代模型，使得控制策略越来越好。

参考文献

1、Yuanlong Li, Yonggang Wen, Kyle Guan, and Dacheng Tao《Transforming Cooling Optimization for Green Data Center via Deep Reinforcement Learning》南洋理工大学

2、Nevena Lazic, Tyler Lu, Craig Boutilier, Moonkyung Ryu《Data center cooling using model-predictive control》Google Research

数据中心能耗预测优化-技术选型

业务建模特点：

- 不太容易打标签（本质为寻优问题，没有最优只有更优，无法人工给数据样本打标签）
- 模型不容易泛化（模型预训练完毕上线后需要不断深化学习跟随env的动态性）
- 安全性要求高（不允许初始模型直接用于环境交互，避免灾难策略动作）
- 参数连续性（控制参数和状态参数都是连续数据）

模型选择： 有范围约束先验知识的强化学习

数据中心能耗预测优化-数据采集

iDCIM系统每隔5分钟将采集到的600多个采集点数据（包括制冷系统、电力系统、环境参数等），上传到AI平台进行数据治理。提取出影响能效指标的20多项关键特征参数用于PUE模型训练。

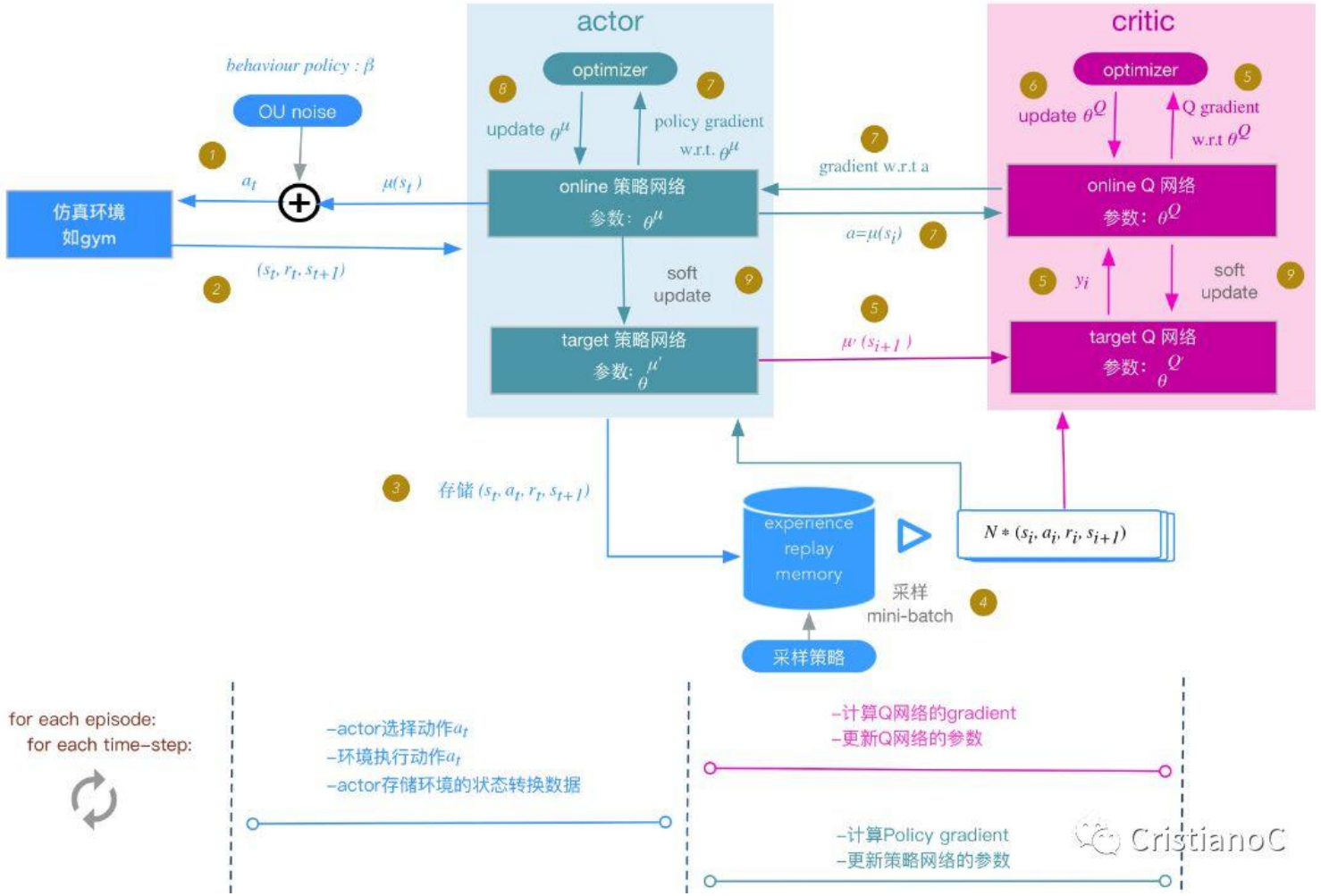
数据中心能耗预测优化-状态表征

水冷系统温度相关项	1	$T_{dec}(t)$	DEC 输出温度
	2	$T_{iec}(t)$	IEC 输出温度
	3	$T_{cw}(t)$	冷冻水管输出水流温度
	4	$T_{dx}(t)$	DX冷冻盘管输出空气流温度
	5	$T_{ch}(t)$	冷却器盘管输出空气流温度
水冷系统物理环节控制相关项	6	$V_{cwwp}(t)$	冷却水水泵PWM状态值
	7	$V_{fwwp}(t)$	冷冻水水泵PWM状态值
	8	$V_{cwvalop}(t)$	冷却水阀门打开量状态值
	9	$V_{fwvalop}(t)$	冷冻水阀门打开量状态值
	10	$V_{cwfan}(t)$	冷却塔风机PWM状态值
	11	$V_{AHUfan}(t)$	AHU风机PWM状态值
水冷系统无关项	12	Tamb(t)	环境温度
	13	Hite(t)	IT载荷
	14	$T_{DCrSet}(t)$	DC room设置温度（DC room的温度要求）
	15	$T_{DCr}(t)$	DC room当前温度
	16	PUE(t)	DC当前的PUE值
	17	DP(t)	AHU进出气管间的气压测

数据中心能耗预测优化-控制策略

1	$SETV_{cwwp}(t)$	冷却水水泵PWM控制量（转速）
2	$SETV_{fwwp}(t)$	冷冻水水泵PWM控制量（转速）
3	$SETV_{cwvalop}(t)$	冷却水阀门打开量
4	$SETV_{fwvalop}(t)$	冷冻水阀门打开量
5	$SETV_{ctfan}(t)$	冷却塔风机PWM控制量（转速）
6	$SETV_{AHUfan}(t)$	AHU风机PWM控制量（转速）

数据中心能耗预测优化-算法原理



图片出处：https://cloud.tencent.com/developer/article/1636750

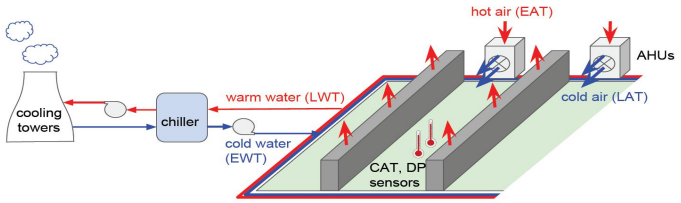
数据中心能耗预测优化-奖励函数

iZCooling奖励函数设计的基本原则是：在保证DC room设定温度要求的前提下降低DC的PUE。

- 参考google和NTU的设计方案，奖励函数设计为：

$$reward = \begin{cases} \gamma \bullet (\dfrac{T_\Phi - T_t}{T_\Phi}), & T_\Phi \le T_t \\ \lambda \bullet (\dfrac{PUE_\Phi - PUE_t}{PUE_\Phi}) + \varphi \bullet (\dfrac{PUE_{t-1} - PUE_t}{PUE_{t-1}}), & T_\Phi \ge T_t \end{cases}$$

数据中心能耗预测优化-方案流程



1	$T_{dec}(t)$	13	Hite(t)
2	$T_{iec}(t)$	14	$T_{DCrSet}(t)$
3	$T_{cw}(t)$	15	$T_{DCr}(t)$
4	$T_{dx}(t)$	16	**PUE(t)**
5	$T_{ch}(t)$	17	DP(t)
6	$V_{cwwp}(t)$		
7	$V_{fwwp}(t)$	1	$SETV_{cwwp}(t)$
8	$V_{cwvalop}(t)$	2	$SETV_{fwwp}(t)$
9	$V_{fwvalop}(t)$	3	$SETV_{cwvalop}(t)$
10	$V_{cwfan}(t)$	4	$SETV_{fwvalop}(t)$
11	$V_{AHUfan}(t)$	5	$SETV_{ctfan}(t)$
12	Tamb(t)	6	$SETV_{AHUfan}(t)$

样本构造 → $(state, \Delta action, reward, \text{state}\_)$

状态转移建模 → $\text{state}\_ = F_1(state, action)$

PUE曲线拟合 → $PUE = F_2(state, action)$

其中：

$$\Delta action = action_i - action_j$$

$$reward = \begin{cases} \gamma \bullet (\frac{T_\Phi - T_t}{T_\Phi}), & T_\Phi \le T_t \\ \lambda \bullet (\frac{PUE_\Phi - PUE_t}{PUE_\Phi}) + \varphi \bullet (\frac{PUE_{t-1} - PUE_t}{PUE_{t-1}}), & T_\Phi \ge T_t \end{cases}$$

第一阶段数据采集：

1.系统状态数据states

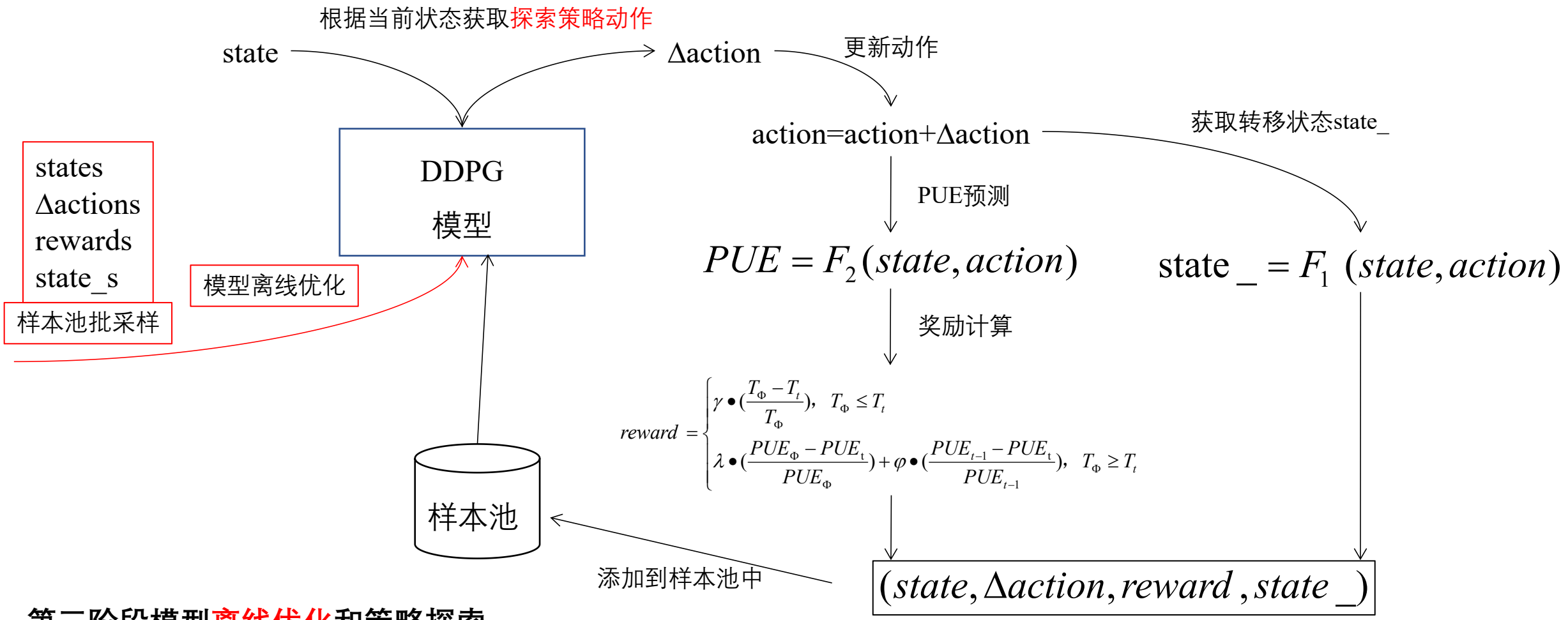
2.控制量的采集actions

第二阶段：

1.PUE、states曲线拟合

2.业务建模，3.样本构建

数据中心能耗预测优化-方案流程



第三阶段模型离线优化和策略探索

数据中心能耗预测优化-方案流程

辅助决策

状态转移模型提升

$state\_ = F_1(state, action)$

PUE曲线拟合提升

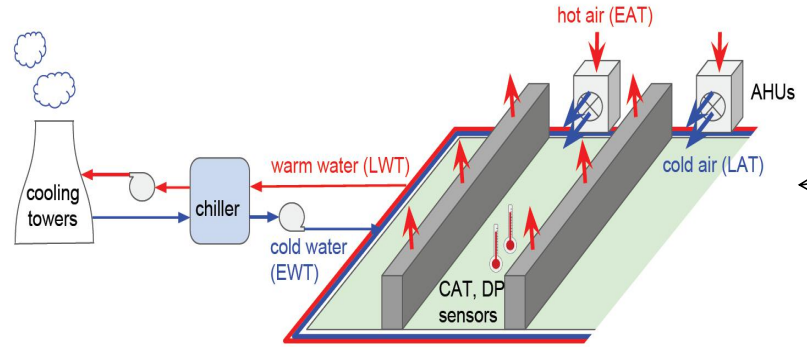
$PUE = F_2(state, action)$

state
action
state_
PUE

模型上线控制

reward
state_

Δaction
action=action+Δaction



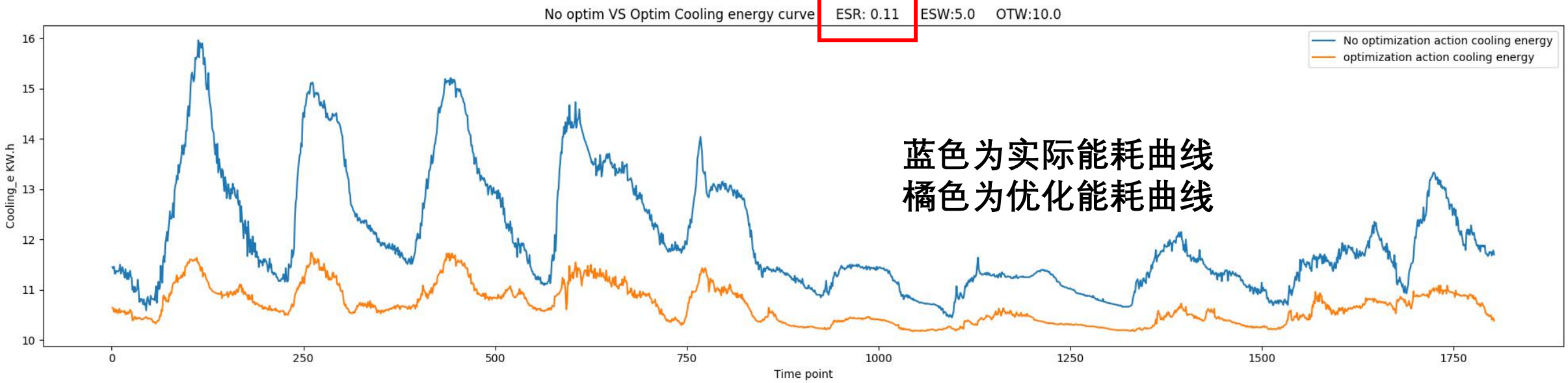
第四阶段：

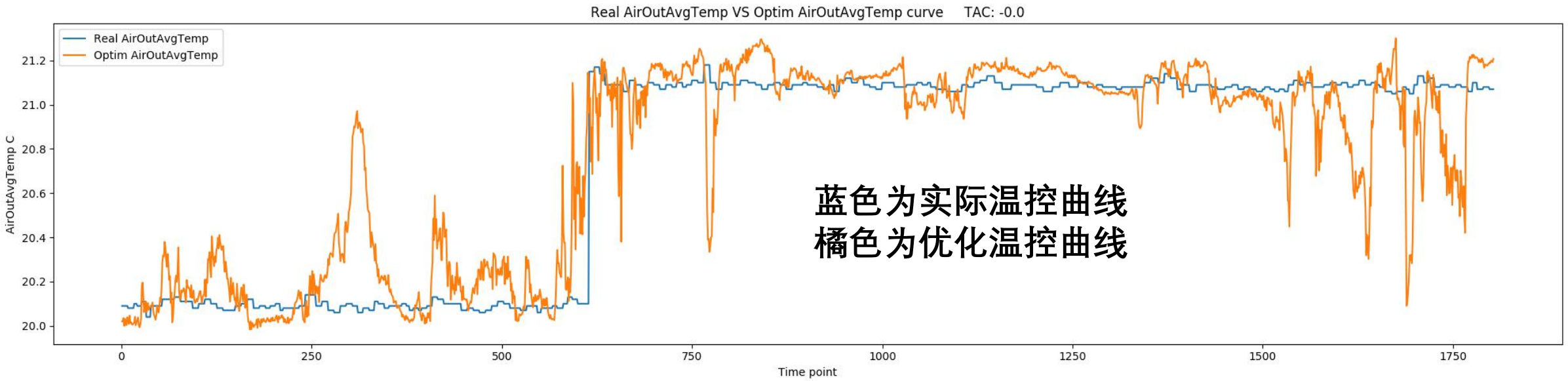
在线模型工作流程及模型优化策略

数据中心能耗预测优化-解决问题

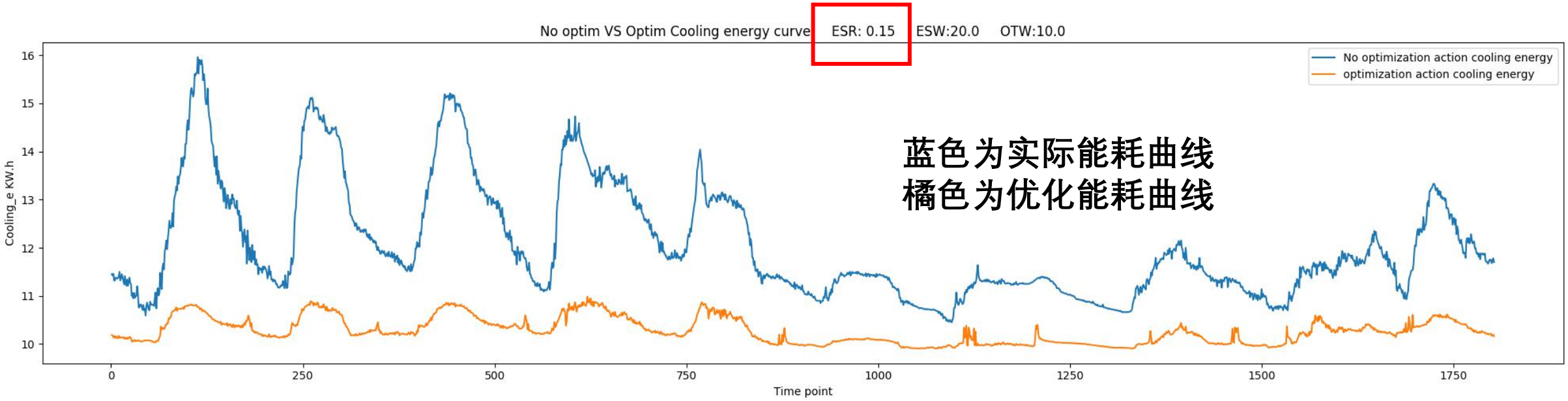
1. 能够解决连续性动作控制问题
2. 能实现模型离线预训练，初始模型无需与环境进行交互，利用历史数据即可优化
3. 对 DC 的水冷系统状态转移和PUE进行建模，增强了控制模型鲁棒性
4. 模型上线后利用交互数据对模型进行提升，适应DC水冷系统在不同季节的波动性
5. 将DC水冷系统的先验知识嵌入到模型中，避免灾难性策动动作

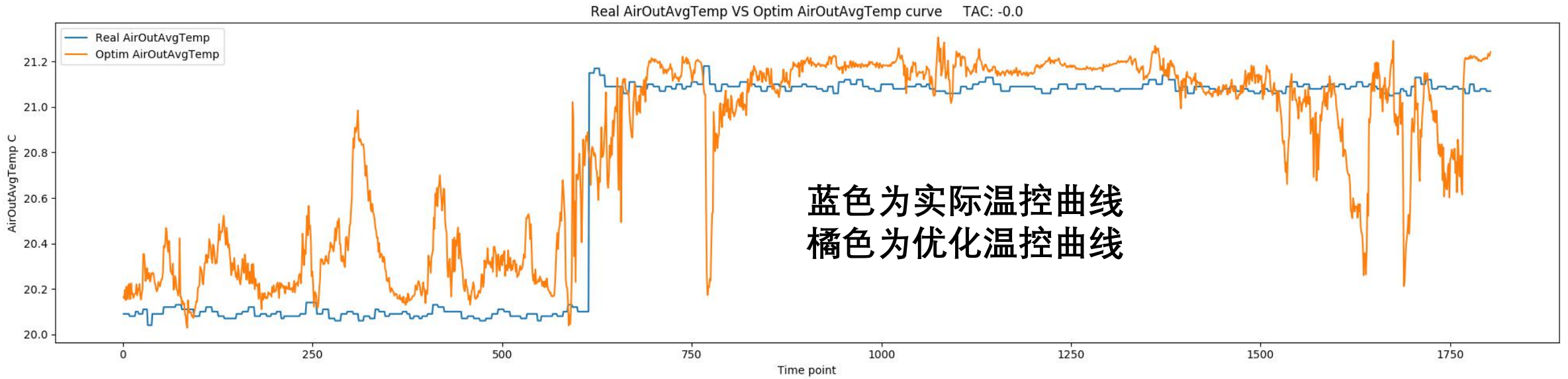
数据中心能耗预测优化-实验结果





数据中心能耗预测优化-实验结果



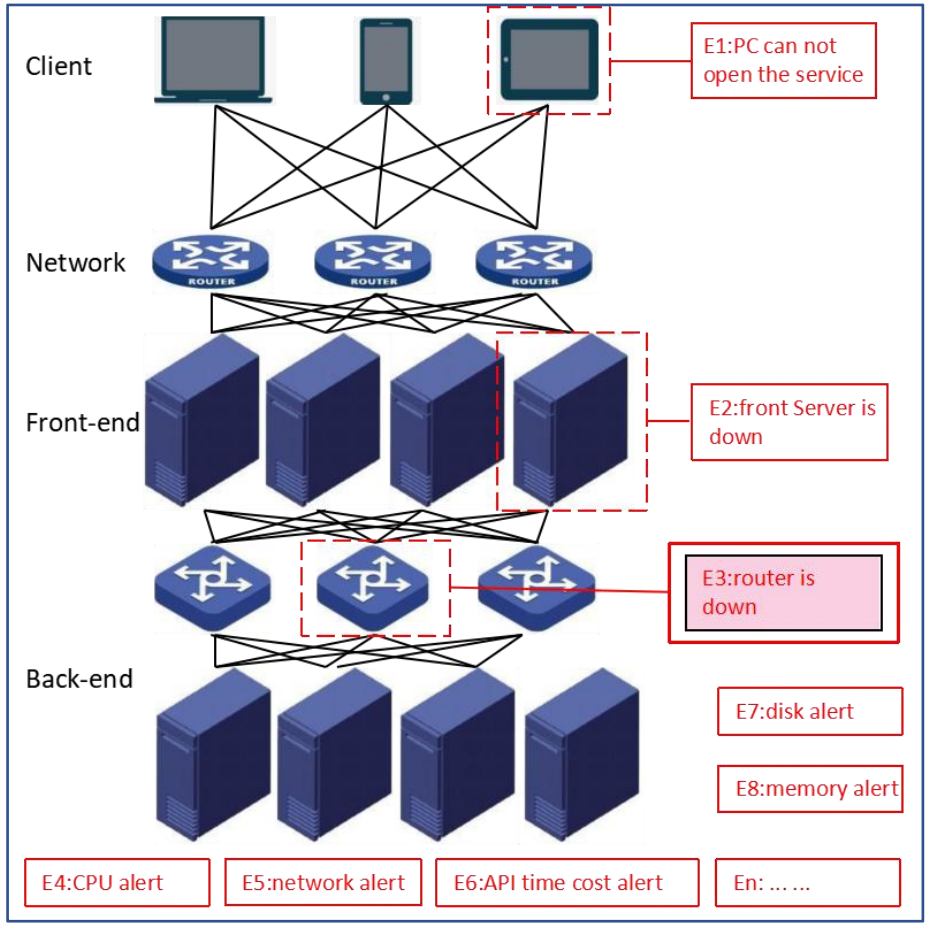


数据中心能耗预测优化-实验结果

实验总结：

实验结果表明在基本满足温控要求的前提下，DCIM优化控制策略能给水冷系统带来**11%~15%**能耗节约。

故障根因智能分析-案例背景



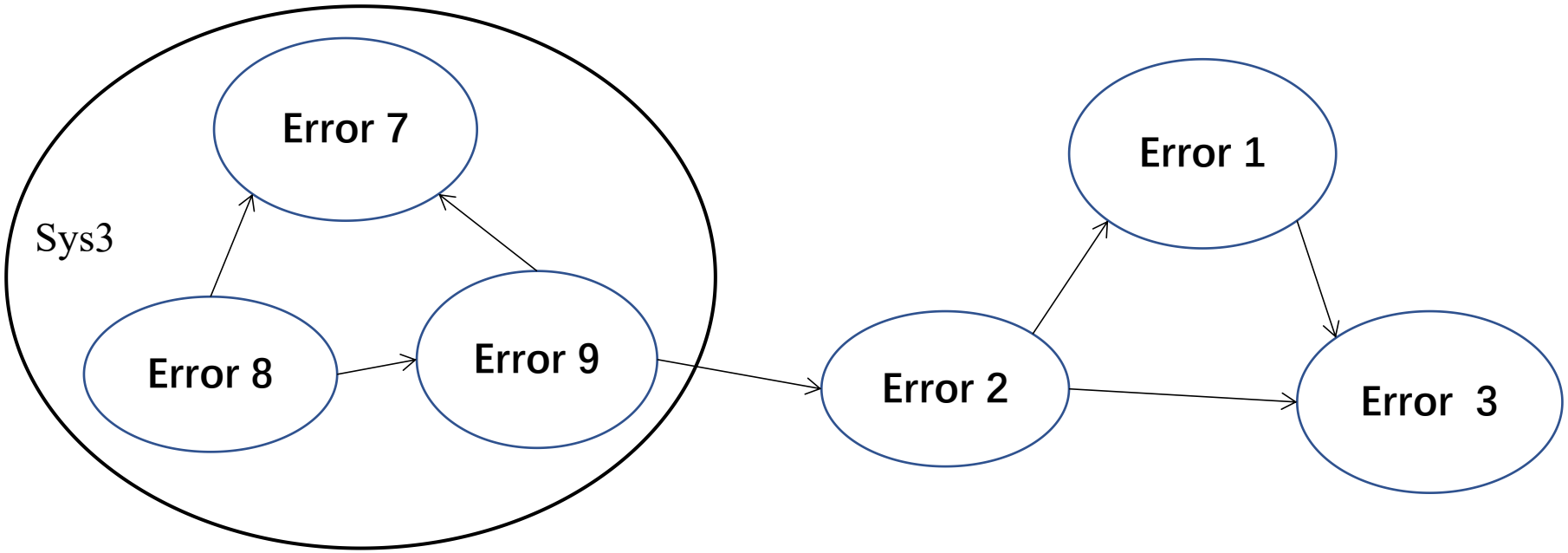
实际场景中路由器故障引发服务器异常示意图

问题： 可以看出到E3路由器发生故障时，系统会产生大量告警信息，导致告警风暴，将有用的告警信息淹没，从而影响故障诊断效率。

根因分析的目标就是在出现大量告警时，对这些告警进行分析处理，过滤掉无效的告警，准确定位出可能的疑似根因节点，缩短故障定位时间，以保证业务的稳定运行。

故障根因智能分析-技术选型

根据系统模块/组件/服务器系统拓扑网络，对系统故障告警Error/Warming信息进行收集并搭建知识图谱，挖掘各告警信息间的因果关系，结合系统网络拓扑的特点，采用图神经网络模型对故障根因进行定位。



故障根因智能分析-技术选型

因为RCA场景涉及到知识图谱、知识推理、网络拓扑的融合，其特点为：

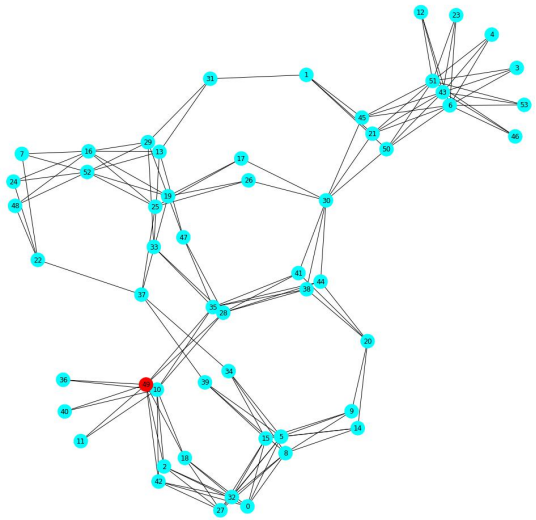
- 非网格化数据，所以无法使用传统的CNN模型
- 数据间距非欧氏距离，所以数据间的距离无法使用物理距离进行表征
- 邻居节点数目不固定

图神经网络技术常用于：节点嵌入、图嵌入、节点分类、关系预测、图分类等场景

模型选择： 基于图神经网络的节点分类

故障根因智能分析-流程框架

系统拓扑信息



网络拓扑图

故障告警信息采集

系统名称	告警时间	告警信息，里面包含ip信息	是否为根因
SYS_8	2019/6/4 1:14	主机node_87 HTTP:http://**********，慢响应（调用耗时超过1000ms）次数：309次（大于阈值：200次）	0
SYS_9	2019/6/4 1:14	主机node_92 上CPU Steal Time持续5分钟超过10%	0
SYS_5	2019/6/4 1:14	主机node_60 端口80通信异常	1
SYS_7	2019/6/4 1:14	主机node_53 HTTP:http://**********，慢响应（调用耗时超过1000ms）次数：309次（大于阈值：200次）	0

节点error日志信息清洗、聚合、特征提取

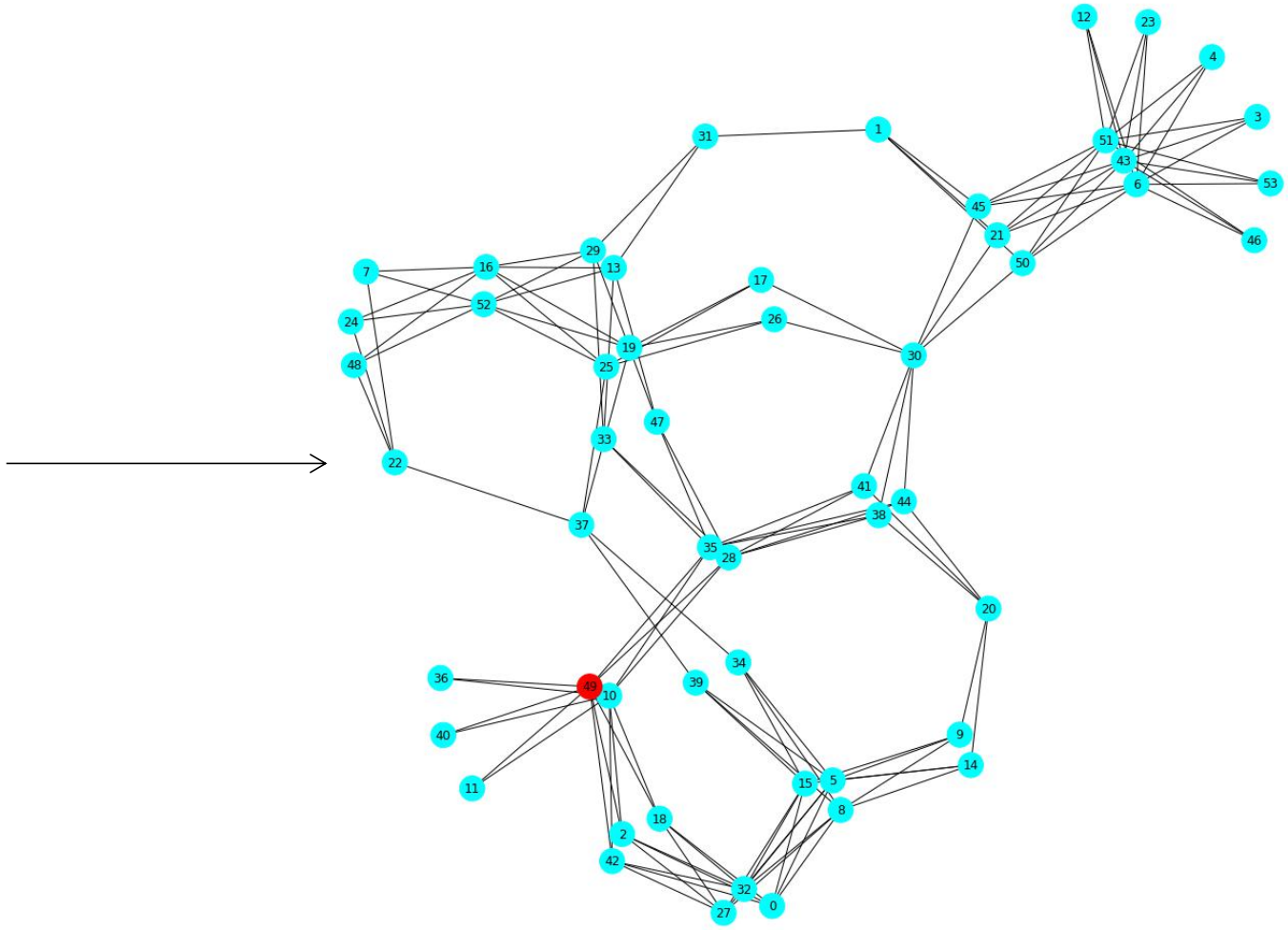
图神经网络推理模型

节点根因判断

故障根因智能分析-拓扑构建

```
{
   "Node1": ["Node4", "Node83", "Node33"],
   "Node2": ["Node4", "Node83", "Node33"],
   "Node3": ["Node4", "Node83", "Node33"],
}
```

其中key值节点为源节点，即调用节点；
value中的节点为目标节点，即被调用节点。



故障根因智能分析-拓扑构建

系统名称	告警时间	告警信息，里面包含ip信息	是否为根因
SYS_8	2019/6/4 1:14	主机node_87 HTTP:http://**********，慢响应（调用耗时超过1000ms）次数：309次（大于阈值：200次）	0
SYS_9	2019/6/4 1:14	主机node_92 #￥%……&*	0
SYS_5	2019/6/4 1:14	主机node_60 端口80通信异常	1
SYS_7	2019/6/4 1:14	主机node_53 HTTP:http://**********，慢响应（调用耗时超过1000ms）次数：309次（大于阈值：200次）	0

故障根因智能分析-信息收敛

系统名称	告警时间	告警信息，里面包含ip信息	是否为根因
SYS_8	2019/6/4 1:14	主机node_87 HTTP:http://**********，慢响应（调用耗时超过1000ms）次数：309次（大于阈值：200次）	0
SYS_5	2019/6/4 1:14	主机node_60 端口80通信异常	1
SYS_7	2019/6/4 1:14	主机node_53 HTTP:http://**********，慢响应（调用耗时超过1000ms）次数：309次（大于阈值：200次）	0

对数字、url、端口、文件路径等字段进行收敛

系统名称	告警时间	告警信息，里面包含ip信息	是否为根因
SYS_8	2019/6/4 1:14	主机node_87 HTTP:*，慢响应（调用耗时超过*ms）次数：*次（大于阈值：*次）	0
SYS_5	2019/6/4 1:14	主机node_60 端口*通信异常	1
SYS_7	2019/6/4 1:14	主机node_53 HTTPS:*，慢响应（调用耗时超过*ms）次数：*次（大于阈值：*次）	0

故障根因智能分析-信息聚合

系统名称	告警时间	告警信息，里面包含ip信息	是否为根因
SYS_8	2019/6/4 1:14	主机node_87 HTTP:*，慢响应（调用耗时超过*ms）次数：*次（大于阈值：*次）	0
SYS_5	2019/6/4 1:14	主机node_60 端口*通信异常	1
SYS_7	2019/6/4 1:14	主机node_53 HTTPS:*，慢响应（调用耗时超过*ms）次数：*次（大于阈值：*次）	0

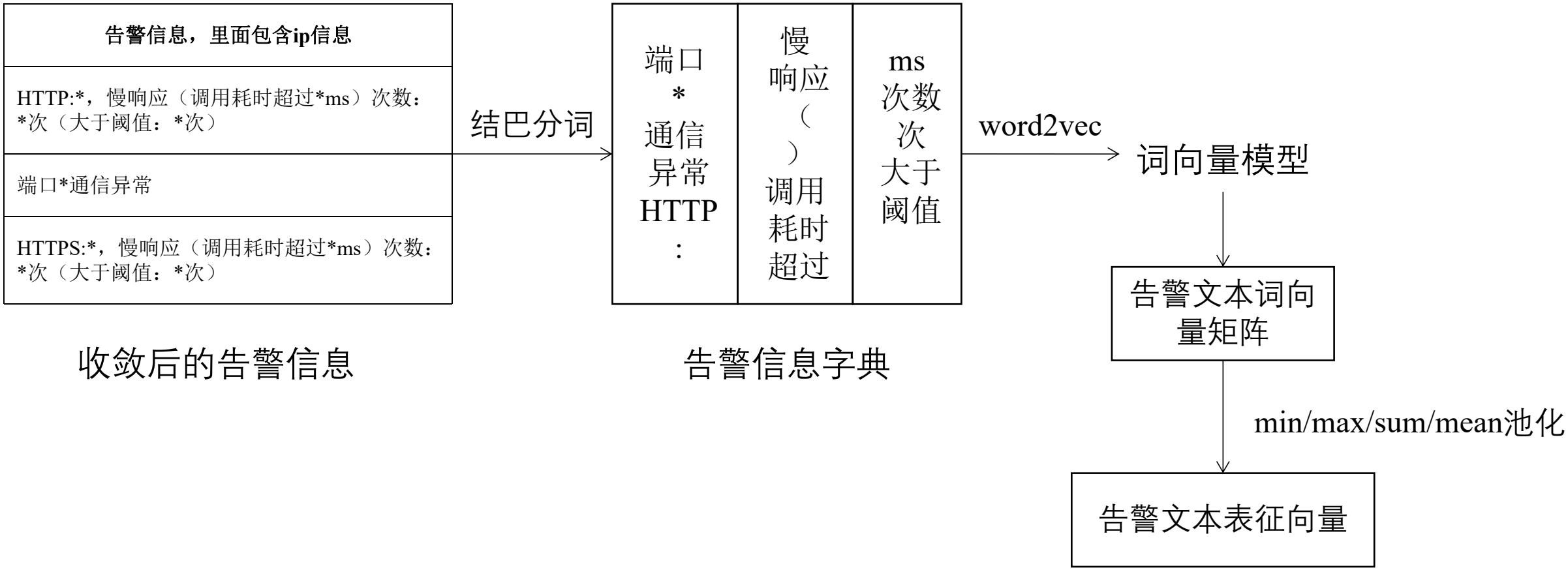
故障根因智能分析-信息向量化

- 向量化方式：根据日志进行分词，使用word2vec获取词向量，通过告警信息词向量池化获取告警信息的表征向量；

优点：表征向量可以表征告警间的相关性，为知识推理提供支撑，能解决未登录告警问题；

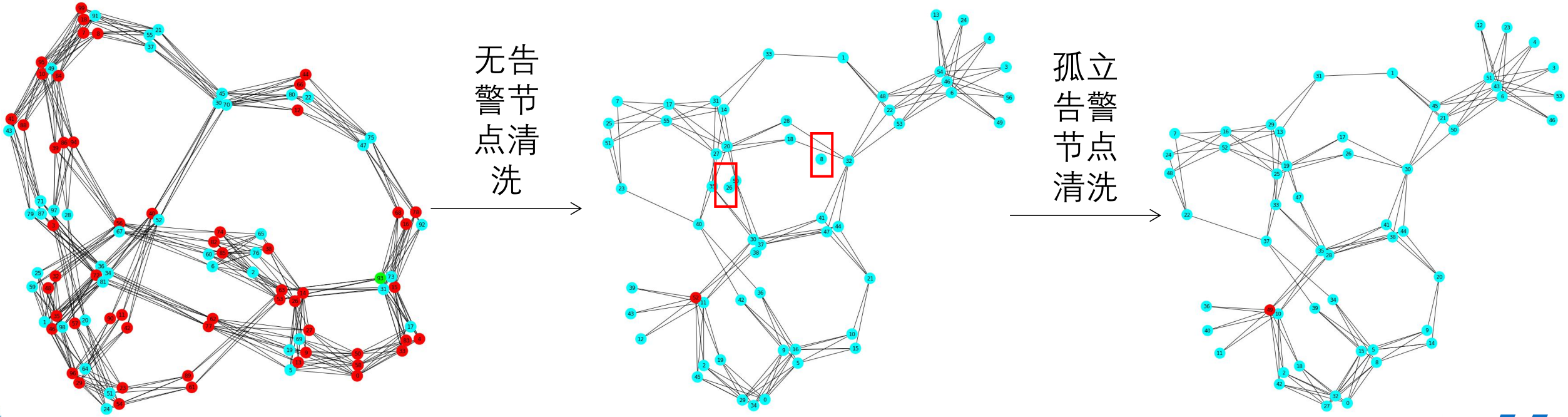
不足：增加模型的计算量

故障根因智能分析-信息向量化



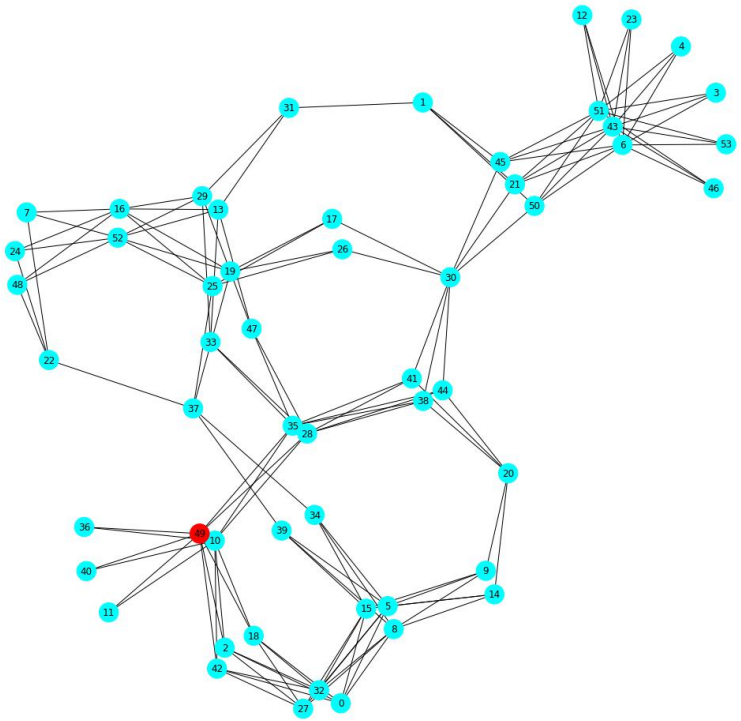
故障根因智能分析-信息向量化

因为故障根因定位的细粒度为节点，而在一次故障发生时并不是所有的节点都会产生告警error/warming信息，或者存在一些告警孤立节点、根据告警风暴的特点和传播效应原理，这些节点通常不会是根因节点，所以需要根据故障的告警信息对系统网络拓扑进行清洗，精简网络拓扑。

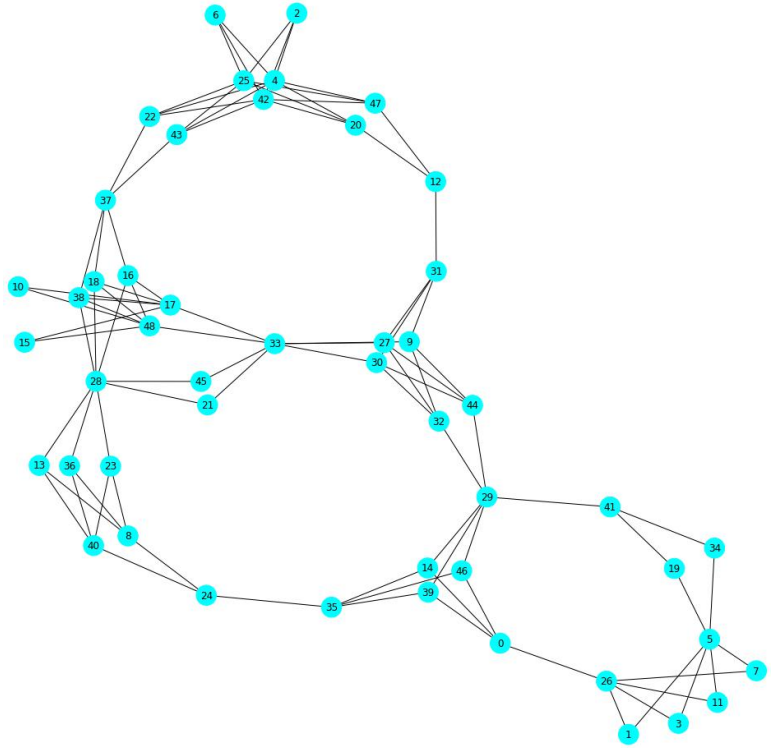


故障根因智能分析-信息向量化

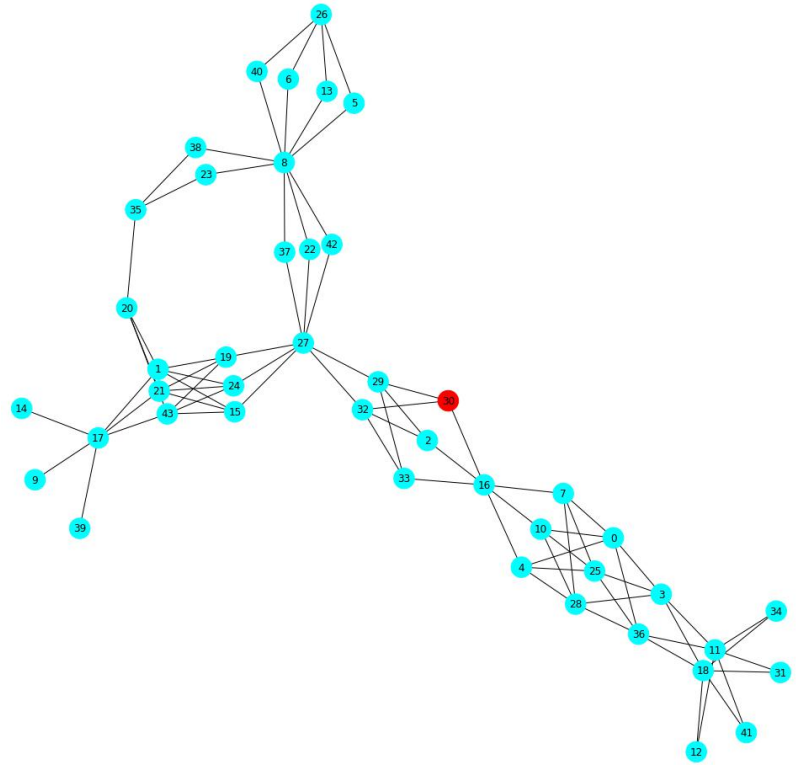
根据以上步骤可以获取每一个故障的告警知识图谱



故障error图谱1（有根因）



故障error图谱2（无根因）



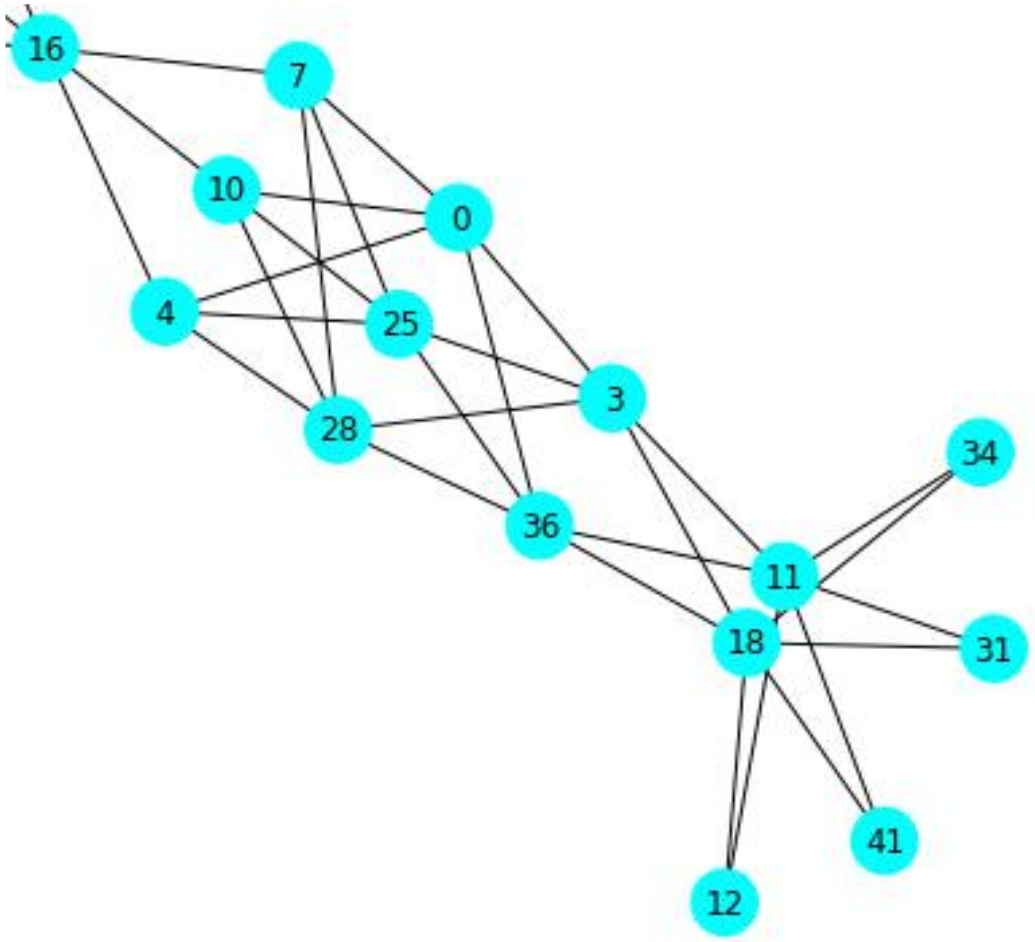
故障error图谱2（有根因）

故障根因智能分析-策略

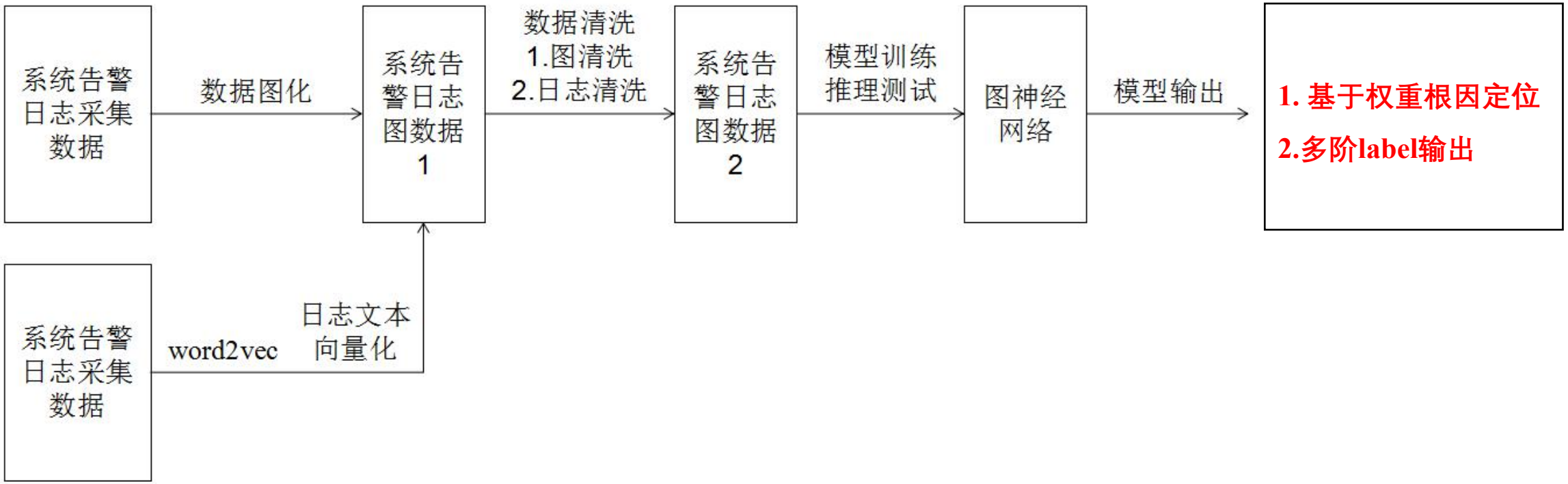
在对一系统进行故障根因定位与图节点分类大体相同，但存在一定的区别:

对于一个故障系统进行根因定位本质上是找出每一个告警节点为根因节点的概率分布，根据概率分布找出最有可能为故障根因的节点。

所以在实际根因定位结果输出的不是直接模型输出的节点分类标签，而是根据网络的softmax层概率分布来输出一阶根因节点、二阶根因节点、三阶根因节点。



故障根因智能分析-整体流程



故障根因智能分析-实验数据

实验数据：苏宁云计算故障根因定位数据集：

训练集100个故障样本，有无根因样本比例为：1:1，每一个样本100个节点

验证集20个故障样本，有无根因样本比例为：1:1，每一个样本100个节点

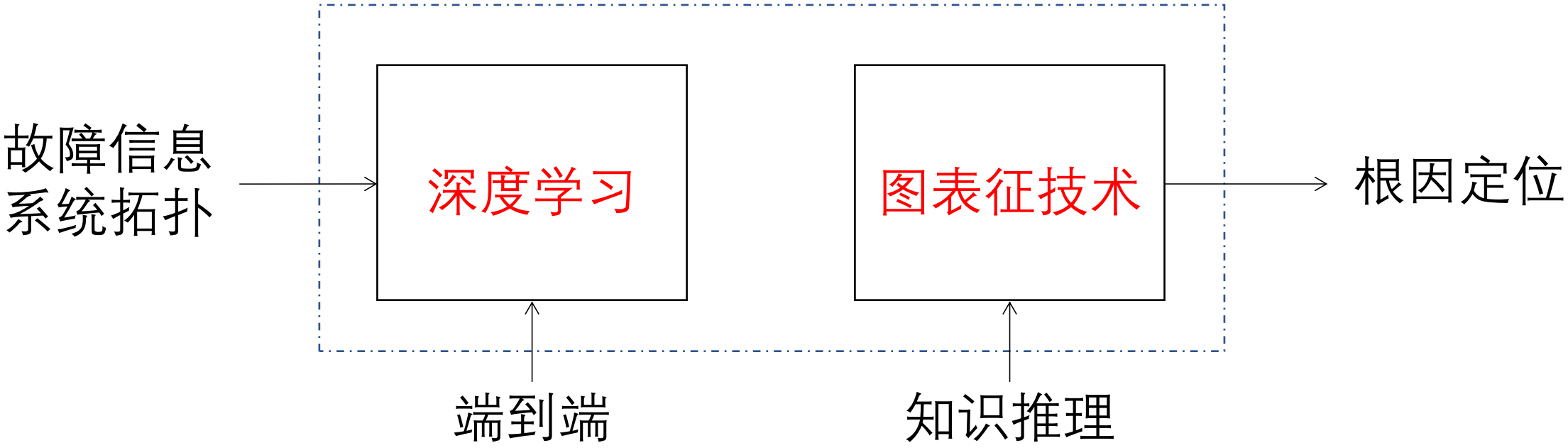
故障根因智能分析-实验结果

msup

	训练集		测试集	
标签阶数	一阶	二阶	一阶	二阶
实验结果1	训练集全样本准确率：98% 训练集正样本召回率：96%	训练集全样本准确率：99% 训练集正样本召回率：98%	测试集全样本准确率：90% 测试集正样本召回率：80%	**测试集全样本准确率：95%** **测试集正样本召回率：90%**
实验结果2	训练集全样本准确率：97% 训练集正样本召回率：94%	训练集全样本准确率：97% 训练集正样本召回率：94%	测试集全样本准确率：95% 测试集正样本召回率：90%	**测试集全样本准确率：100%** **测试集正样本召回率：100%**
实验结果3	训练集全样本准确率：99% 训练集正样本召回率：98%	训练集全样本准确率：100% 训练集正样本召回率：100%	测试集全样本准确率：85% 测试集正样本召回率：70%	**测试集全样本准确率：95%** **测试集正样本召回率：90%**
实验结果4	训练集全样本准确率：97% 训练集正样本召回率：94%	训练集全样本准确率：98% 训练集正样本召回率：96%	测试集全样本准确率：90% 测试集正样本召回率：80%	**测试集全样本准确率：95%** **测试集正样本召回率：90%**
实验结果5	训练集全样本准确率：97% 训练集正样本召回率：94%	训练集全样本准确率：98% 训练集正样本召回率：96%	测试集全样本准确率：85% 测试集正样本召回率：70%	**测试集全样本准确率：95%** **测试集正样本召回率：90%**

故障根因智能分析-总结

结合深度学习&图表征技术能快速、有效地、实现系统端到端的故障根因定位，我们的实验证明一阶故障根因定位正确率>80%，二阶故障根因定位正确率>90%，为后续故障自恢复提供有效支撑。



故障根因智能分析-下一步计划

- 对于根因节点无告警日志的场景做研究
- 多于多故障的根因分析场景做研究

关注msup公众号

获取更多AI落地实践

麦思博(msup)有限公司是一家面向技术型企业的培训咨询机构，携手2000余位中外客座导师，服务于技术团队的能力提升、软件工程效能和产品创新迭代，超过3000余家企业续约学习，是科技领域占有率第1的客座导师品牌，msup以整合全球领先经验实践为己任，为中国产业快速发展提供智库。