南开大学本科生创新科研计划项目

实施计划书

项目类型： 南开大学创新科研百项工程项目

项目编号： 202210055828

项目名称： 大规模无线基站多维指标异常检测系统设计与实现

项目成员： 张怡桢、潘骁腾、张家冉

指导教师： 张圣林

填表日期： 年 月 日

说明

1.“实施计划书”是项目立项之后,根据批准项目的类别(“国创”、“市创”、“百项”)不同的资助额度和执行时限，结合评审答辩老师的意见和建议。对申请项目进行必要的调整，并制定实施方案和安排

一、实施计划书:

项目实施方案（项目研究目标、研究内容、技术路线、方法等）：

研究内容与目标

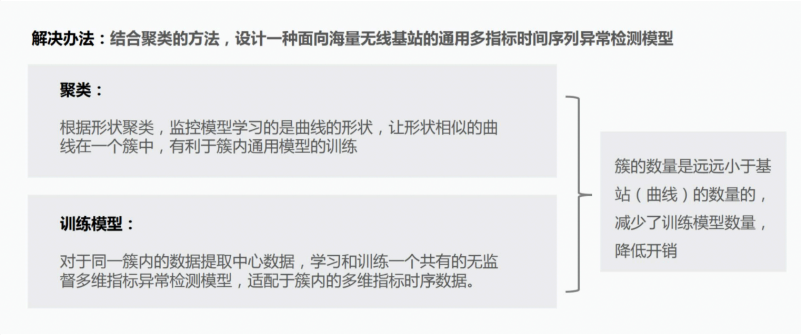
针对基于传统机器学习算法的多维指标时间序列异常检测方法无法表达无线基站监测数据复杂 时序性和关系性，以及基于深度学习算法的多维指标时间序列异常检测方法需要较长周期的训练 数据和巨大的训练开销的问题，设计并实现新型面向大规模无线基站的多维指标时间序列异常检 测系统。首先拟研究海量的多维指标数据基于形状的方法，使得同一类内的多维指标数据形状相 似。然后对于同一类内的数据提取中心数据，学习和训练一个共有的无监督多维指标异常检测模 型，适配于类内的多维指标时序数据。最后优化模型的时效性，实现大规模无线基站多维指标时 序数据异常检测系统，服务中国移动公司的无线网络监测场景。

技术路线

多维指标异常检测的基本思想是首先使用GRU捕捉原始多维指标数据间的复杂时间依赖关系，并结合一种常用的算法——VAE将输入变量映射到随机隐藏变量。其次，为了明确地建立随机 隐藏空间中随机变量之间的时间依赖关系，提出一种随机变量连接技术：使用线性高斯状态空间 模型（SSM）连接随机变量，并将随机隐藏变量与GRU隐藏变量进行拼接。最后，为了帮助变分 网络中的随机变量捕获输入数据的复杂分布，采用planar NF学习随机隐藏空间中的非高斯后验分 布。

模型训练模块的目的是学习捕获多维指标数据的正常模式并为每个数据点输出异常分数。阈值 选择模块使用离线训练时的异常分数，根据POT（Peaks-Over-Threshold）方法自动选择异常阈值 。离线训练模块可以定期执行（例如每周或每月进行一次）,从而适应最新的数据特征。

根据已经训练好的模型，预处理后的新数据点将输入到在线检测模块，得到其异常得分。如果 该时刻的异常得分高于异常阈值，则将该时刻判断为异常，否则为正常。如果检测某时刻发生异 常，通过评估该时刻每个指标的贡献（即重建概率）来解释这次异常。



项目实施进度和安排：

项目实施进度和安排		
阶段	进度安排	具体安排
第一阶段：2022年3月—8月	学习项目需求的预备知识	（1）基础的分析方法：可视化分析、相关分析、信息熵增益分析等。 （2）学习机器学习相关的统计学习模型与算法知识：逻辑回归算法、关联关系挖掘、聚类算法、决策树算法、随机森林算法、支持向量机模型、蒙特卡洛树搜索算法、隐式马尔可夫模型、多示例学习算法、迁移学习算法、卷积神经网络算法等。 （3）了解智能运维：在有编程语言、数学、机器学习算法的基础上，学习智能运维的相关知识，了解掌握智能运维体系的基本架构，为后面的实现打下基础。 （4）学习Linux Shell脚本语言：掌握常用Linux Shell命令，熟练使用Linux操作系统。
第二阶段：2022年9-10月	收集整理互联网公司的真实数据	收集部分互联网公司服务器的KPI数据（CPU利用率、每秒查询的数量、响应延迟、PV、GC等运维数据）。所收集的数据应包括已经标记了异常指标的数据和未标记的原始数据，然后将得到的数据进行进一步整理、筛选和分类，为下一步的预处理操作做准备。
2022年11月—2023年2月	预处理、聚类、提取特征值	（1） 对事先收集的不同类型的数据进行标准化，将它们的数据缩小到一个大致相同的范围，使其具有可比性。 （2） 通过聚类算法，将大量的数据分成几个集群，并分别处理每个集群，得到对应集群中心的KPI曲线。 （3） 将每种特征参数化，从已经标记过的数据的集群中心提取KPI曲线的特征值，对于新的数据，把该KPI曲线的特征值、所属集群中心的特征值和异常标记作为训练集。
第四阶段：2023年3月-4月	实现服务指标异常检测模型	尝试采用算法OmniAnomaly(1/2)和CTF(Coarse-to-Fine Model Transfer)(2/2)作为半监督学习的算法来检测KPI曲线的异常。基于已形成的服务异常指标检测模型，可以实现自动对一组新的KPI曲线进行异常检测。将各个特征量化，通过其对应的阈值来判断是否存在异常情况。
第五阶段：2023年5月—7月	服务指标异常检测模型的训练及改进	将不断出现的新的KPI曲线分配到一个已经存在的集群中，合并新的KPI曲线中未标注的数据和聚类中心，最后通过半监督学习训练成一个新的KPI曲线模型，实现服务指标异常检测模型的不断更新，确保其判断的准确率能得到进一步的提升。
第六阶段：2023年8月—11月	数据的再收集及正确性的验证	重新收集在阶段二的采样企业近十个月的数据，并与服务指标异常检测模型得到的结果进行对比，测试模型的正确性。
第七阶段：2023年12月—2024年2月	模型评估	（1）真实数据采集：从知名互联网公司收集真实的原始运维数据。 （2）对于每种特征，设定阈值，并选定评估异常检测的度量方法。 （3）将本项目所采取的方法实际的性能与基于监督学习方法和非监督学习方法的性能建立一定的度量比较方法进行比较，分别从准确率、数据需求量、人力劳动资源的需求量等角度进行评估比较。

预期成果（中期成果、最终成果）：

中期成果：希望得到一个初步的检测模型。

最终成果：将人工智能技术引入无线网络运维，可以为运维场景 提供高准确性和及时性的服务，降低人力成本，提高运维效率，并提升用户满意度，检测异常的模型准确率预期达到80%。

拟开展研究性学习交流活动（活动主题、活动的目的和意义、活动内容、参加对象、活动形式）：

活动主题：OmniCluster拓展组会

活动的目的和意义：将人工智能技术引入无线网络运维，可以为运维场景 提供高准确性和及时性的服务，降低人力成本，提高运维效率，并提升用户满意度，检测异常的模型准确率预期达到80%。

活动内容：通过交流与会议确定项目的各步进展，项目的进行方向与方法，通过研读论文来确定聚类方法与选择的机器学习算法，针对“ 如何解决基于传统机器学习算法的多维指标时间序列异常检测方法无法准确表达无线基站监测数 据复杂时序性和关系性，以及基于深度学习算法的多维指标时间序列异常检测方法需要较长周期的训练数据和巨大的训练开销 ”等问题提出解决方案。

参加对象：张家冉，张怡桢，潘晓腾，张圣林老师，李正丹老师

活动形式：线上会议

二、经费预算:

预算科目	支出项目	金额
实验业务费	云服务平台租用测试费	1000 元
实验材料费	计算机配件，元器件购置费	600 元
图书资料费	购书费，复印费	200 元
其他	宣传，调研费用	200 元
合计		2000 元

三、审批情况:

指导教师意见：

年　　月　　日