南开大学本科生创新科研计划

项目申请书

〇国家级大学生创新训练计划

〇天津市大学生创新训练计划 /百项工程

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

负 责 人： 张怡桢

所在学院： 软件学院

联系电话： 18060766269

指导教师： 张圣林

填表日期： 2022年03月20日

填表说明

一、申请书逐项认真填写，填写内容必须实事求是，表达明确严谨。

二、“所属学科”按一级学科列出，

三、“起止时间”写到项目主持人毕业时间以前

四、“项目申请人（团队）知识背景、创新经历、特长、兴趣；取得成果”要提供证明材料，附在表后，证明可以是复印件。

五、“指导教师推荐意见”要对申报项目内容进行全面分析。

六、材料规格：用A4纸双面打印（复印），左侧装订。

一、基本信息

研究项目	**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			
	项目性质	应用性研究			
	项目来源	教师指导选题	**所属学科**	软件工程	
	起止时间	年　月　至　年　月			
申请者		**项目组长**	**成员1**	**成员2**	**成员3**
	姓名	张怡桢	潘骁腾	张家冉	
	学号	2013747	2013132	2012685	
	专业	软件工程	软件工程	软件工程	
	所在院系	软件学院	软件学院	软件学院	
	电话	18060766269	13908455948	19898931612	
	微信号	jryzxf	qhd1904	zxn19898931612	
指导教师	**姓名**	**职称、职务**	**单位**	**电话**	**E-mail**
	张圣林		软件学院	18801349816	zhangsl@nankai.edu.cn
	李正丹	初级实验师	软件学院	19922573135	lzd@nankai.edu.cn

项目申请人承诺：

我保证申请材料的真实性。如果获得立项，严格遵守实事求是的原则，恪守学术道德规范，认真实施项目计划，积极开展研究性学习和交流活动，合理支出项目经费，保证完成实验目标。 签名：**张怡桢 潘骁腾 张家冉**

二、项目实施方案:

1. 立项依据与研究内容(研究意义、国内外研究现状及发展动态分析）:

随着4G/5G无线网络在人们生活中的重要性日益增长，用户对无线网络服务质量的要求越来越高，给网络运维增加了难度和压力。传统的运维模式以较多人力参与为主，运维效率以及准确性已无法满足需求。在当前降本增效提质的压力下，急需引入人工智能方法来为运维赋能。

近年来，以机器学习和深度学习为代表的人工智能算法在无线网络异常检测、关联分析、根因定位等关键问题上取得了突破性的进展。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。

为了持续监测无线基站的健康状态，运维人员为每个无线基站配置了监测指标，并不断采集这些监测指标的数据。因此，一个指标的监测数据形成一个单指标时间序列，具有多个指标的无线基站监测数据形成一个多维指标时间序列。无线基站多维指标时间序列的异常行为往往表征了无线基站出现了异常并预示潜在的故障。因此，如果能及时检测多维指标时间序列的异常，就能迅速发现无线基站的异常，从而避免无线基站故障带来的损失。

由于无线网络场景复杂多变，运维数据海量且关联关系复杂，因此当前基于机器学习或者深度学习的异常检测方法不能很好的适用于无线网络。具体而言，当前基于传统机器学习的异常检测算法虽然可以解决海量数据问题，但对多维指标的复杂时序性和关系性表达能力不强。基于深度学习的异常检测算法虽然可以很好的表达多维指标时序数据，但是它们往往需要较长周期的训练数据和巨大的训练开销。如何设计一种适用于海量无线基站多维指标时间序列数据的异常检测算法，是一个亟待解决的挑战性问题。

2. 项目的研究内容、研究目标、技术路线:

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。

3. 拟解决的关键问题和实现方法:

关键问题:

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的问题，是本项目需要解决的关键问题。

实现方法：

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。

4. 本项目的特色与创新之处:

创新性：

本项目拟使用随机变量连接以挖掘模型的潜在表示，创新性地从分布更复杂、历史周期更长的数据中学习多维指标的正常模式。本项目拟通过结合基于聚类+异常检测的算法，在保证较高检测准确度的前提下，大大降低了模型的训练开销。算法将应用于中国移动无线网络的运维场景，解决多维指标时间序列的聚类和异常检测问题。

研究内容将面向中国移动大规模网络中的运维场景进行强针对性的算法研发。算法方案除了具有一定普适性外，将对中国移动特定运维场景进行调整与适配，以期能在中国移动大规模复杂网络中具有良好的效果。

5. 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。

6. 主要参考文献:

Zhihan Li, Youjian Zhao, Rong Liu, Dan Pei. Robust and Rapid Clustering of kpis for Large-Scale Anomaly Detection. IEEE/ACM IWQoS 2018, Ban , Alberta, Canada, June 4-6, 2018 (CCF 推荐 B 类会议)

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, Dan Pei. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. ACM SIGKDD 2019, Anchorage, Alaska, USA, August 4-8 2019 (CCF 推荐 A 类会议)

7. 项目申请人（团队）创新经历:

张怡桢：南开大学软件学院2020级本科生，具有一定的编程基础与能力，对计算机网络，数据库原理和应用，java语言基础和工程应用，机器学习等具有浓厚兴趣。

潘骁腾：南开大学软件学院2020级本科生，有C++基础，自学Python，选修MATLAB与经典统计学相关课程。

张家冉：南开大学软件学院2020级本科生，有一定的C++编程基础，自学Python相关课程，对机器学习、计算机操作系统有浓厚的兴趣.

8. 预期成果:

2021-2023年，预期完成“ 无线网多维指标异常检测算法 ”研究内容的方案，拟研究针对海量多维指标数据的聚类算法，使得同一簇内的数据形状较为相似；然后对于每个簇内的数据提取簇中心，训练对应的无监督多维指标异常检测模型，适配于簇内的多维指标数据，从而实现无线网多时间序列聚类及异常检测的通用性算法研究 ；最后优化模型的时效性，保证模型可以具有多维指标时序数据异常检测的实用性，服务于无线网络在线监测场景。

9.项目简介（200字以内）：

随着网络的发展，人们对无线网络服务质量的要求越来越高，及时检测并发现无线基站的异常，从而避免无线基站故障带来的损失成为了一个重要的研究命题。设计一种适用于海量无线基站多维指标时间序列数据的异常检测算法，在传统机器学习算法和深度学习算法中从中找到一个平衡的解决办法，是本项目的研究重点。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。

10.项目英文简介:

With the development of network, people have higher and higher requirements on the quality of wireless network service. It has become an important research proposition to detect and discover the abnormality of wireless base station in time so as to avoid the loss caused by the failure of wireless base station. The research focus of this project is to design an anomaly detection algorithm suitable for multi-dimensional index time series data of massive wireless base stations and find a balanced solution between traditional machine learning algorithm and deep learning algorithm. The introduction of artificial intelligence technology into wireless network operation and maintenance can provide high accuracy and timely services for operation and maintenance scenarios, reduce labor costs, improve operation and maintenance efficiency, and improve user satisfaction.

三、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	5000 元
实验材料费	计算机配件，元器件购置费	3000 元
图书资料费	购书费，复印费	1000 元
其他	宣传，调研费用	1000 元
合计		10000 元

四、审批情况:

指导教师推荐意见：

此项目研究难度适中，有利于培养学生的科研能力。

2022年03月21日

学院评审意见：

年　　月　　日

学校主管部门审批意见：

年　　月　　日