南开大学本科生创新科研计划项目

中期报告书

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

负 责 人： 张怡桢

指导教师： 张圣林

填表日期： 2022年11月13日

一、基本信息

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	
项目组成员	学号	个人完成情况及对项目的贡献
张怡桢	2013747	数据预处理，数据清洗
潘骁腾	2013132	将数据集进行聚类
张家冉	2012685	数据异常标注，清洗数据
指导教师	李正丹,张圣林	

二、中期检查报告

1. 项目进展情况（包括已完成研究目标和内容、开展的研究性学习和交流）：

1. 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。

2. 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。

3. 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Mc2PCA聚类算法，将数据集聚类成几个集群。并对聚成的簇的结果与已经标注好的标签比较，进行聚类指标评估。

2. 阶段性成果:

1. 完成知识储备
2. 收集整理互联网公司的真实数据，即字节数据集和移动数据集
3. 将N个MTS对象的MTS数据集X划分到K个集群（簇C）中
4. 得到聚类的结果，即输出簇和每个簇的成员
5. 对聚类结果用指标进行评估

3. 项目进行中存在的问题（包括未按原计划实施的原因）及解决办法:

1. 训练数据过大，导致模型无法运行出结果。首先减少了数据集的条数，尝试跑通模型，再进行下一步优化。然后陆续更换了数据集和标注，评估结果有所提升。
2. 评估聚类效果的指标中，有时NMI指标非常差但ACC指标较高，即模型结果很差。除了调整模型本来的参数外，还对原始数据集进行了预处理的工作（即对每个实例进行标准化之后再输入模型）。实验得出预处理后的评估指标结果有所提升。
3. 为了做出较好的聚类效果，多次尝试聚成不同个数的簇，即每次训练更改K的值。这里K=5、10、20、50、100均有尝试。
4. 采用该模型进行聚类的结果总是只有一个簇。原因可能是数据没有标准化，没有标准化的数据计算协方差矩阵时，容易产生较大的差异，造成聚类结果崩塌。

4. 中期之后的工作计划:

1. 分别处理聚类之后的每个集群，得到对应集群中心的KPI曲线。

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。

2. 实现服务指标异常检测模型

尝试采用算法OmniAnomaly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。

3. 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。

三、审批情况

指导教师意见： 年　　月　　日
学院评审意见和建议： 年　　月　　日
学校主管部门审批意见： 年　　月　　日