2021 年中国高校大数据挑战赛

赛题 A：智能运维中的异常检测与趋势预测

异常检测（异常诊断/发现）、异常预测、趋势预测，是智能运维中首当其冲需要解决的问题。这类问题是通过业务、系统、产品直接关联的 KPI 业务指标进行分析诊断，指标主要包括用户感知类（如页面打开延时）、服务性能（如用户点击量）、服务器硬件健康状况（如 CPU 利用率、内存使用率）等关键性能指标。

不同场景的运维，分析的指标种类差异较大，但都具备时序性特点，不同场景的 KPI 指标，以毫秒、秒、分钟、小时、天为时间间隔的数据序列都会出现，有些复杂场景的业务，往往会混合多个时间间隔的数据，但均为随时间变化而变化的时序数据。

本次赛题以运营商基站 KPI 的性能指标为研究数据，数据是从 2021 年 8 月 28 日 0 时至 9 月 25 日 23 时共 29 天 5 个基站覆盖的 58 个小区对应的 67 个 KPI 指标。其中，选取三个核心指标进行分析。

第一个指标：小区内的平均用户数，表示某基站覆盖的小区一定时间内通过手机在线的平均用户人数；

第二个指标：小区 PDCP 流量，通过小区 PDCP 层所发送的下行数据的总吞吐量(比特)与 小区 PDCP 层所接收到的上行数据的总吞吐量(比特)两个指标求和得到，表示某基站覆盖的小区在一定时间内的上下行流量总和；

第三个指标：平均激活用户数，表示某基站覆盖的小区在一定时间内曾经注册过无线网络的平均人数。

针对上面三个指标，完成如下 3 个问题：

问题 1 异常检测：利用附件的指标数据，对所有小区在上述三个关键指标上检测出这 29 天内共有多少个异常数值，其中异常数值包含以下两种情况：异常孤立点、异常周期。

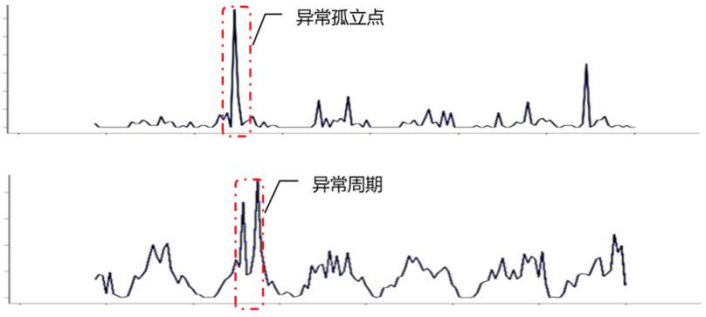

图 1 两种网元（小区）异常情况

汇总所有小区的异常情况填写如下表格。

表 1：异常检测汇总表

	时间周期选择标准	异常孤立点的个数	异常周期个数
小区内的平均用户数			
小区 PDCP 流量			
平均激活用户数			

问题 2 异常预测：针对问题 1 检测出的异常数值，通过该异常数值前的数据建立预测模型，预测未来是否会发生异常数值。异常预测模型除了考虑模型准确率以外，还需要考虑两点：1）模型输入的时间跨度，输入数据的时间跨度越长，即输入数据量越多，模型越复杂，会增加计算成本和模型鲁棒性，降低泛化能力；2）模型输出时间跨度，即预测的时长，如果只能精准预测下一个时刻是否发生异常，在时效性上则只能提前一个小时，时效性上较弱。

问题 3 趋势预测:利用 2021 年 8 月 28 日 0 时至 9 月 25 日 23 时已有的数据，预测未来三天（即 9 月 26 日 0 时-9 月 28 日 23 时）上述三个指标的取值。并完整填写附件 2 中的预测值表格，单独上传到竞赛平台。

说明:

（1）异常孤立点，在一段时间内仅有 1 个异常值；异常周期，在一段时间内有多个异常值。(时间范围、异常值范围需要参赛者自行设定并说明理由)。

（2）在异常预测和趋势预测时，可借用其他指标作为辅助输入特征建模，如预测第 i 个指标第 t 时刻的数值或是否异常，可使用第 j 个指标 t 时刻前的数值作为输入，但不能以第 j 个指标第 t 时刻及之后的数值作为输入。

（3）异常预测和趋势预测建模时，需考虑每个小区、基站之间的差异。可以针对每个小区、基站单独建模，也可以统一建模，最终以模型评价指标来评估。

（4）第一题和第二题以 F1 值（2*精确率*召回率/（精确率+召回率））来评估模型优劣，第三题以 MAPE（平均绝对百分比误差）作为模型评估指标。