2021 年中国高校大数据挑战赛 A 题思路

赛题 A：智能运维中的异常检测与趋势预测

供参考思路

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。

针对问题 1:该问需要找到所有小区中上述三个指标的异常数值。根据提示的异常数值的两种情况。可以得知，该问题的需要分别对三个指标进行分析，分别找出每个指标的异常点与周期。（如果使用综合考虑三个指标，然后利用 LOC、COF、孤立森林等可以利用多个变量进行异常值检测的方法，可能就不太符合题意了）。因此可以终点考虑与时间序列相关的异常值发现方法。可以先针对序列，可视化出每个指标的波动情况，观察数据趋势后，再进一步的对数据进行分析，找到异常值。查找时序异常值的一种方式，就是利用你和模型的置信区间来约束其取值（注意：三个指标的取值应该均为大于等于 0 的数值）。

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。

该问题需要：针对第一问异常值检测结果，建议一个二分类模型（判断后面的一个时刻或者几个是逗是否出现异常值）。建议先从一后一个时刻是否会出现异常点入手。这对这样的情况，就首先是数据的特征工程处理，即需要解决使用哪些特征作为模型的输出，使用前面几个时间段的数据用于是否为异常值模型的预测，在特征工程将数据处理好之后，可以使用一些合适的机器学习模型建立分类模型进行预测。

在预测后一个小时的模型建立好之后，再考虑利用相同的方式，建立更有实效性的模型。

特征工程常用的方式：数据降维（PCA、KPCA、tSNE）、特征选择、特征组合（特征之间相互组合得到新的特征）。
机器学习的分类模型：SVM、随机森林、神经网络、与时序相关的 LSTM 等。

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。

针对问题 3：该问题需要针对已经有的数据，预测为了三天的数据。针对单个指标与单个基站和小区，可以看作为是一个单变量的时间序列数据的建模与预测。当然针对该问题，可以一为单个指标或者多个指标统一建立时序的预测模型。多个指标的综合考虑的建模与预测，会使问题更加的复杂。但是该问与时序有很强的关系，不可以忽略时间的作用。
时间序列常用模型：ARIMA、SARIMA、prophet、指数平滑，等算法。