**C题 数据驱动的异常检测与预警问题**

推动生产企业高质量发展，最根本的底线是保证安全、防范风险，而生产过程中产生的数据能够实时反映潜在的风险。附件1为某生产企业某日00:00:00-22:59:59由生产区域的仪器设备记录的时间序列数据（已经进行数据脱敏），本题未给出数据的具体名称，这些数据可能是温度、浓度、压力等与安全密切相关的数据。

请依据附件1数据，建立数学模型，完成以下问题：

**问题1：**附件1所给出的数据都可能存在波动，且所有波动都在安全值范围内。有些波动可能是正常性波动，例如随着外界温度或者产量变化的波动，或者可能是传感器误报，这些波动具有规律性、独立性、偶发性等特点，并不能产生安全风险，我们视为非风险性异常，不需要人为干预；有些波动具有持续性、联动性等特点，这些异常性波动的出现是生产过程中的不稳定因素造成的，预示着可能存在安全隐患，我们视为风险性异常，需要人为干预、分析和评定风险等级。请建立数学模型，给出**判定非风险性异常数据和风险性异常数据**的方法。【分类】

**【数据处理：**100个传感器，不知道数据名称的情况，必然是要进行变量筛选的。首推计算100列之间的欧氏距离做聚类再筛选。这里的数据降维不要进行标准化、归一化处理，把每一类中选择几个作为新数据集就可以了。】【非风险性异常：单独列中有独立点的存在，只需使用简单的离群点检测即可，并将其剔除。】

**【**对于每列先找出不在正常范围内的数据（假设所有数据服从正向分布，在某个值外面的认为是不正常数据）做上标记，然后根据某一时间点所有传感器数据横向，如果在阈值个数之上则认为是纷现形异常。】

**问题2：**结合问题1的结果，建立数学模型，给出**风险性异常数据**异常程度的量化评价方法，要求使用百分制（0-100分）对每个时刻数据异常程度进行评价（分值越高表示异常程度越高）。应用所建立的模型和附件1的数据，找到数据中异常分值最高的5个时刻及这5个时刻对应的异常传感器编号（每个时刻只填风异常的。写5个异常程度最高的传感器编号，异常传感器不足5个则无需填满；如果得分为0，可以不用填写异常传感器编号），并给出数学模型对所得结果进行评价。

表1 问题2的结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 第一高分 | 第二高分 | 第三高分 | 第四高分 | 第五高分 |
| 异常程度得分 | 2461 | 2443 | 2431 | 2429 | 2392 |
| 异常时刻编号 | 9:58:15 | 9:58:30 | 9:58：30 | 9:55：45 | 9:55：15 |
| 异常传感器编号 |  |  |  |  |  |
| 异常传感器编号 |  |  |  |  |  |
| 异常传感器编号 |  |  |  |  |  |
| 异常传感器编号 |  |  |  |  |  |
| 异常传感器编号 |  |  |  |  |  |

**问题3：**为了提前发现未来生产过程中可能存在的风险隐患，请建立风险性异常预警模型，预测当日23:00:00-23:59:59可能产生的风险性异常。结合问题2中给出的风险性异常程度量化评价方法，指出23:00:00-23:59:59中四个时间段（见表2），每个时间段内的最高异常分值及对应的异常传感器编号（只填写5个异常程度最高的传感器编号，异常传感器不足5个则无需填满；如果得分为0，可以不用填写异常传感器编号）。

表2 问题3的结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 时间 | 23:00:00-23:14:59 | 23:15:00-23:29:59 | 23:30:00-23:44:59 | 23:45:00-23:59:59 |
| 异常最高分 |  |  |  |  |
| 异常传感器编号 |  |  |  |  |
| 异常传感器编号 |  |  |  |  |
| 异常传感器编号 |  |  |  |  |
| 异常传感器编号 |  |  |  |  |
| 异常传感器编号 |  |  |  |  |

**问题4：**根据问题2和问题3中的结果，建立数学模型对该生产企业整个生产系统的安全性进行评价，请在00:00:00-23:59:59内每隔30分钟，用0-100分进行安全性评分，0分表示安全性最低，100分表示安全性最高（包括00:00:00-23:00:00的得分和23:00:00-23:59:59的预测得分），并用适当的方法对所给评分的结果进行评价和敏感性分析。

表3 问题4的结果

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 安全性评分 | 时间 | 安全性评分 |
| 00:30:00 |  | 12:30:00 |  |
| 01:00:00 |  | 13:00:00 |  |
| 01:30:00 |  | 13:30:00 |  |
| 02:00:00 |  | 14:00:00 |  |
| 02:30:00 |  | 14:30:00 |  |
| 03:00:00 |  | 15:00:00 |  |
| 03:30:00 |  | 15:30:00 |  |
| 04:00:00 |  | 16:00:00 |  |
| 04:30:00 |  | 16:30:00 |  |
| 05:00:00 |  | 17:00:00 |  |
| 05:30:00 |  | 17:30:00 |  |
| 06:00:00 |  | 18:00:00 |  |
| 06:30:00 |  | 18:30:00 |  |
| 07:00:00 |  | 19:00:00 |  |
| 07:30:00 |  | 19:30:00 |  |
| 08:00:00 |  | 20:00:00 |  |
| 08:30:00 |  | 20:30:00 |  |
| 09:00:00 |  | 21:00:00 |  |
| 09:30:00 |  | 21:30:00 |  |
| 10:00:00 |  | 22:00:00 |  |
| 10:30:00 |  | 22:30:00 |  |
| 11:00:00 |  | 23:00:00 |  |
| 11:30:00 |  | 23:30:00 |  |
| 12:00:00 |  | 23:59:59 |  |

完成时间为：7月31日至8月6日20点，每个队只提交1份，由队长提交，将论文电子版发至：[mathchen@126.com](mailto:mathchen@126.com)

要求：1文件名培训队号+队长名；

2 论文正文5号字，标题4号字，字体不限，正文不少于15页(不包括摘要，目录及附录)；

3本次练习老队员可以不提交论文，新队员根据论文提交及论文质量原则上末位淘汰不超过15队（<10%），如果总体论文质量较高，可以不淘汰。