2022 第二届天府杯全国大学生数学建模竞赛论文

题　　目　　仪器故障智能诊断技术

摘　要：

本文主要通过分析仪器设备产生的波形信号数据，提取重要信号特征，进而达到对设备仪器的故障智能检测。根据信号数据的特点，我们对问题 1 绘制箱线图对异常值进行降噪处理；对问题 2 用主成分分析法提取主成分特征，再用随机森林算法提取重要特征；对问题 3 用 K-均值聚类算法、高斯混合分布进行设备故障的分类；对问题 4 用随机森林及其优化、高斯朴素贝叶斯的方法进行设备故障的分类；对问题 5 用控制变量法解决“提取特征数量以及质量对特征识别的影响”的问题。

对于问题 1，我们新添 3 个指标分别为 R-Square、MAE、MAPE。首先建立“箱线图异常点调整模型”。然后用 PyCharm 搭建 Python3.8 编译环境，计算每一列数值的箱线图的上边缘值和下边缘值。将每一列所有异常点调整至上、下边缘值之间。最后，用 Python 绘制每一列数据调整前和调整后的箱线图进行比较，分析 6 项指标以验证模型降噪的合理性，并将处理后的结果汇总在附表 1-1 和 1-2 中。

对于问题 2，我们首先读取降噪后的数据，在已搭配环境下运用主成分分析法建立“主成分特征提取模型”。在对该模型的基础之上运用随机森林算法建立“重要性特征提取模型”。在对两个模型合理的理论证明和推导后，从“主成分特征提取模型”中得到 72 个主成分特征，再从“重要性特征提取模型”中提取更为重要的 16 个特征，最终得到维度为 100×16 的样本特征数据，并将其结果汇总在附表 2-1 和 2-2 中。

对于问题 3，我们新添 3 个指标分别为精确率、F1-score、错误率。读取问题 2 得到的样本特征数据，在已搭建环境下构建“K-均值聚类算法模型”，预测准确率均值为 58%，但未能达到题意要求。再根据高斯分步函数建立“混合高斯模型”。最终得到样本预测准确率均值为 91%，准确率标准差为 0.29。

对于问题 4，我们用问题 2 中提取到的样本特征数据，在表中给两类故障分别加上标签。首先，在已搭配环境下用多种分类器给数据进行分类，得到最高分类准确率分别是随机森林分类器和高斯朴素贝叶斯分类器，且准确率都为 90%。在此基础之上我们运用控制变量法以及模型得分准则对参数进行调整，运用特殊的训练集和测试集划分进行改进，进而建立 “随机森林模型”和 “高斯朴素贝叶斯模型”。最终得到“随机森林模型”中分类器的准确率为 93.33%，预测准确率均值为 90%；“高斯朴素贝叶斯模型” 中分类器的准确率为 95%，预测准确率均值为 97%。

对于问题 5，我们利用主成分分析法提取贡献率累计达到 95%的 72 个特征，利用 100 个样本数据的 72 个特征在随进森林和高斯朴素贝叶斯分类器中进行分类，得出分类准确率分别为 83.33%和 86.67%。再利用随进森林分类器对 100 个样本进行单一特征分类，从而选取最佳特征组合，得到最终分类结果为 90%。

关键词　箱线图、主成分分析法、随机森林、混合高斯、高斯朴素贝叶斯

一、问题重述

仪器设备故障诊断技术是一种了解和掌握机器在运行过程的状态，确定其整体或局部正常或异常，早期发现故障及其原因，并能预报故障发展趋势的技术。随着计算机技术和人工智能科学的发展，基于机器学习或深度学习的故障智能诊断方法成为从业者的新型决策工具，其中故障类型识别技术特点在于：降低原始数据的环境噪声或异常数据影响，提取可靠的波形特征判据，选择或改进现有的机器学习方法，设计一系列必要的仿真实验，讨论与分析。

问题 1 要求我们对附件一和附件二进行降噪处理，并新添三项指标和给定指标对降噪效果进行评价。

问题 2 要求我们对附件一和附件二分别提取“重要信号特征”，以能够作为故障智能检测。并将结果分别保存在附表 2-1 和 2-2 中。

问题 3 要求我们基于无监督或者半监督的方法对附表数据进行二分类，并通过新添三项指标和给定指标保证预测方法的预测准确率均值在 90%以上，准确率标准差在 10 以内。最后，将结果保存在附表 3。

问题 4 要求我们基于有监督方法对附表数据进行二分类，并通过新添三项指标和给定指标保证预测方法的预测准确率均值在 95%以上，准确率标准差在 5 以内。最后，将结果保存在附表 4。

问题 5 要求我们三选一解答问题，我们选择“讨论特征的数量以及质量对特征识别的影响，以及给出解决办法”。

二、问题分析

（一）问题 1 的分析

问题 1 要求我们对附件数据进行降噪处理，新添 3 项指标。降噪目的是找出噪声并减少或排除噪声对后续工作的影响。分析 6 项指标对降噪效果进行评价，以检验模型降噪的合理性。

问题 1 属于数据预处理问题。读取原始数据，首先建立“箱线图异常点调整模型”。计算每一列数值的箱线图的上边缘值和下边缘值。将每一列数值大于该列上边缘值的异常点调整为上边缘值，并将每一列数值小于该列下边缘值的异常点调整为下边缘值。最后，绘制数据调整前和调整后的箱线图进行比较，并通过 6 项评估该模型的合理性。

（二）问题 2 的分析

问题 2 属于提取主要特征的数学问题，要求我们在降低数据维度的同时要提取最重要的特征。

问题 2 一般的数据降维分析方法有奇异值分解、主成分分析、因子分析以及独立成分分析等。由于题中所给数据为 4096 列信号数据，而主成分分析法适合对信号数据进行处理。因此，我们先建立“主成分特征提取模型”，然后在该模型基础上再建立“重要性特征提取模型”。最后，将这两个模型提取的样本特征数据分别在多种分类器中进行二分类，通过分类准确率验证特征模型的合理性。

（三）问题 3 的分析

问题 3 要求我们运用无监督学习方法对两类数据进行分类并新添 3 项指标。

问题 3 属于无标签分类问题，读取问题 2 的样本特征数据。我们先构建“K-均值聚类算法模型”，然后通过高斯分步概率密度函数建立“混合高斯模型”。将每一个未测试的样本作为测试集，将其余 70%样本作为模型训练集，30 作为测试集。最后记录 6 项指标。最后，通过比较两个模型分类的准确率，得到最终符合要求的模型。

（四）问题 4 的分析

问题 4 要求我们运用有监督学习方法对两类数据进行分类并新添 3 项指标。

问题 4 属于分类问题，读取问题 2 的样本特征数据，在表中给两类故障分别加上标签。首先，通过多种分类器以及参数调整优化建立“随机森林分类器模型”，和 “高斯朴素贝叶斯模型”。最后，通过比较两个模型分类的准确率，得到最终符合要求的模型。

（五）问题 5 的分析

问题 5 要求我们讨论并解决特征的数量以及质量对特征识别的影响。

对于问题 5，我们利用主成分分析法提取贡献率累计较高的多个特征，利用 100 个样本数据的这些特征在随进森林和高斯朴素贝叶斯分类器中进行分类，得出分类准确率。再利用随进森林分类器对 100 个样本进行单一特征分类，从而选取最佳特征组合，可得得到最终分类结果。

三、模型假设

1. 假设题目所给的数据真实可靠；
2. 假设不同仪器设备之间运行时不会相互影响；
3. 假设所有仪器设备信号数据都在相同运行时间段截取；
4. 假设所有仪器设备都在同一环境下运行；
5. 假设所有仪器设备使用年限一样。

四、定义与符号说明

序号	符号	说明
1	UE	箱线图的上边缘
2	LE	箱线图的下边缘
3	Q_3	箱线图的上四分位数
4	Q_2	箱线图的中位数
5	Q_1	箱线图的下四分位数
6	IQR	箱线图的四分位距
7	x_i	第i个样本（i=1, 2, ⋯, 100）
8	w_j	投影后的新坐标
9	$\mathrm{Info}(D)$	随机变量 D 的信息熵
10	$\mathrm{Info}(D\|A)$	特征 A 对随机变量 D 的条件熵
11	$Gain(D\|A)$	特征 A 对随机变量 D 的信息增益

12	$\mathrm{GainRate}(D\|A)$	特征 A 对随机变量 D 的信息增益率
13	$\mathrm{Gini}(p)$	基尼系数
14	$\mathrm{Gini}_A(D)$	特征 A 对随机变量 D 的基尼系数
15	$P(x\|\theta)$	高斯分布概率密度
16	$P(y_i\|x)$	朴素贝叶斯条件概率

五、模型的建立与求解

第一部分：准备工作

（一）数据的处理

1. 数据完整性分析

读取 100 个 txt 样本数据，最后整合得到维度为 100×4096 的数据矩阵，故该数据没有缺失值。

2. 数据周期性分析

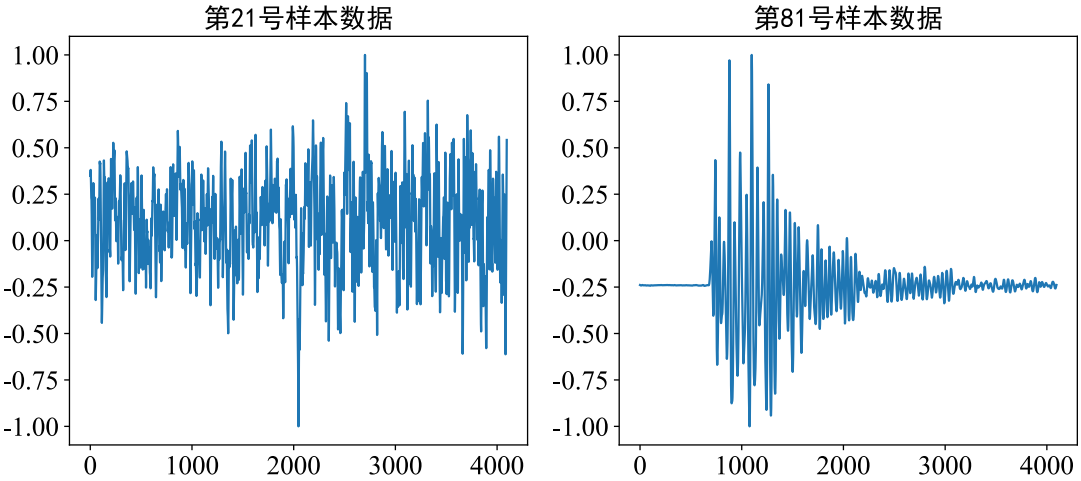


图 5-1 第 21 和 81 号样本信号图

从附件 1 和附件 2 中分别（随机）取一个样本，第 21 和 81 号样本（随机取）信号数据图如图 5-1 所示，难以发现其数据周期性。绘制每个样本的数据图，也难以发现数据存在周期性。

（二）预测的准备工作

我们从原始数据中随机读取 19 列数据，如图 5-2 部分降噪前数据箱线图所示，可知有少部分噪声。因此需要对数据进行降噪处理。

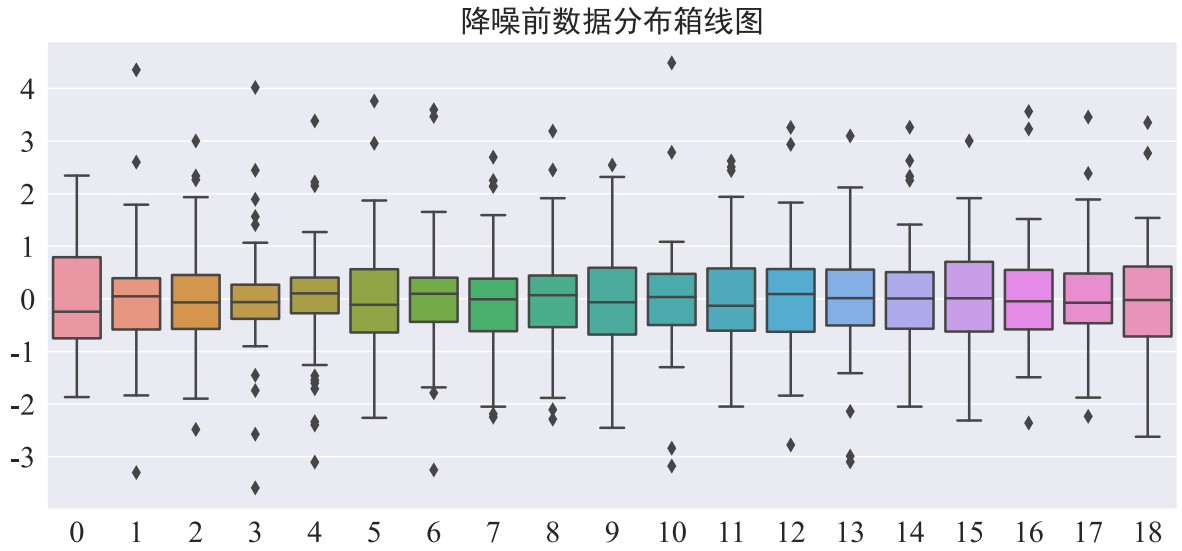
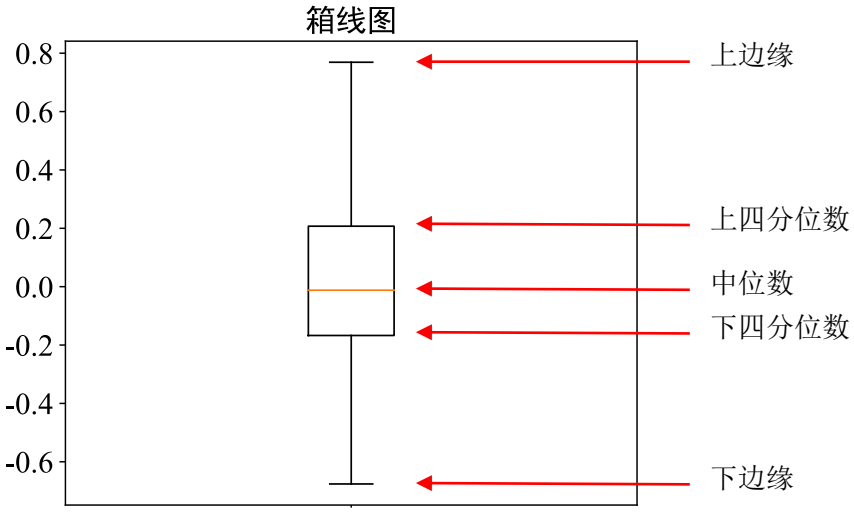


图 5-2 部分降噪前数据

第二部分：问题 1 的一个模型

（一） 箱线图异常点调整模型



5-3 箱线图及其位点

1. 箱线图理论

箱线图[1]四分位数的含义是，一组数据按照从小到大顺序排列后，把该组数据四等分的数，称为四分位数。下四分位数Q_1、第二四分位数Q_2（也称“中位数”）和上四分位数Q_3分别等于该样本中所有数值由小到大排列后第 25%、第 50%和第 75%的数字。上四分位数与下四分位数的差距称为四分位距IQR。上边缘计算公式：

$$UE = Q_3 + \frac{3}{2}IQR \tag{5.1}$$

下边缘计算公式：

$$LE = Q_1 - \frac{3}{2}IQR \tag{5.2}$$

2. 箱线图异常点调整模型的建立与求解

读取 100 个 txt 文件并将其整合，得到维度为 100×4096 的数据矩阵。考虑每一列

信号的特征，对每一列数据进行异常值调整。通过计算箱线图的上边缘和下边缘，把小于下边缘对应的数值改为下边缘对应值，把大于上边缘对应的数值改为上边缘对应值。其异常点调整前后对比如下图 5-4（以其中某列为例）。

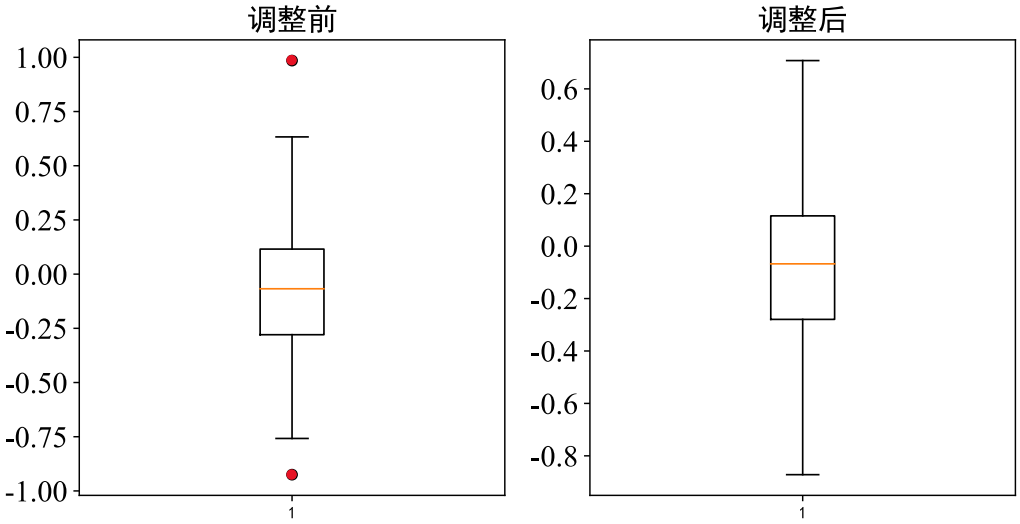


图 5-4 箱线图异常值调整前和调整后的比较

如图 5-2 所示，调整前在超出上边缘和下边缘界限的异常点（红色点）各有一个，调整后异常点消失。其降噪效果六项指标 MSE、SSE、RMSE、R-Square、MAE、MAPE 部分结果如下表 1 和 2 所示，其结果保存在文件中“附表 1-1.csv”和“附表 1-2.csv”中。其中 R-Square、MAE、MAPE 为新添指标。

表 1 附件 1 的数据降噪效果

样本序号	MSE	SSE	RMSE	R-Square	MAE	MAPE
1	0.00055	2.25453	0.02346	0.99450	0.00638	0.00391
2	0.00842	34.49529	0.09177	0.91984	0.05944	0.02911
…	…	…	…	…	…	…
70	0.00035	1.43524	0.01872	0.99462	0.00351	0.00199

表 2 附件 2 的数据降噪效果

样本序号	MSE	SSE	RMSE	R-Square	MAE	MAPE
1	0.00005	0.22474	0.00741	0.99803	0.00051	0.00042
2	0.00016	0.63769	0.01248	0.99537	0.00106	0.00075
…	…	…	…	…	…	…
30	0.00014	0.57497	0.01185	0.99676	0.00126	0.00090

由表 1 和 2 可知，MSE、RMSE、MAPE、MAE 都非常小，符合实际情况；SSE 有部分数值较大，也符合实际情况，因为存在噪声影响较大的点；观察 R-Square 的数值，只有 2 号样本为 0.919844，其余都大于 0.962356，接近 1，表明效果较好。综合上述 6 项指标，箱线图异常点调整模型降噪效果非常好。

第三部分：问题 2 的两个模型

（一）主成分特征提取模型

1. 主成分分析法理论[2][3]

主成分分析法通过投影达到降低数据维度目的。具体过程包括对所有样本进行中心化、计算样本协方差矩阵、对协方差矩阵做特征向量分解、取最大的d^l个特征值对应的特征向量。中心化计算公式：

$$\sum_{i=1}^{100} x_i = 0 \tag{5.3}$$

设$X = \{x_1, x_2, \dots, x_{4096}\}$，则协方差矩阵可表示为$XX^T$，设新坐标系为$W = \{w_1, w_2, \dots, w_d\}$，则协方差矩阵做特征向量分解公式为：

$$XX^TW = \lambda W \tag{5,4}$$

将特征值排序，再去前d^l个特征值对应的特征向量构成的$W = (w_1, w_2, \dots, w_{d^l})$，其中$d^l < d$，$W$即为主成分分析的解。

2. 主成分特征提取模型的建立与求解

根据主成分分析法，从 4096 列数据中取 100 列。综合考虑特征的数量和主成分方差累计贡献率，经过两者权衡，将其方差贡献率按从大到小排序。结果表明贡献率低的特征有 30 个左右，并且计算排序最后的 28 个特征的贡献率总和不到 5%。最终决定主成分方差累计贡献率取 95%，所得主成分共 72 个，满足后续分类算法的要求，其部分结果如下表 3 所示，完整数据保存在附件“AandB_feature_NO_lable_data.csv”。

表 3 72 个主成分特征

样本序号	指标 1	指标 2	…	指标 71	指标 72
1	0.69055	1.47229	…	0.72993	-0.52326
…	…	…	…	…	…
50	-0.57530	-1.87110	…	-0.78037	-0.86305
…	…	…	…	…	…
100	-0.25604	0.20059	…	0.03309	-0.98530

（二）重要性特征提取模型

1. 随机森林理论[4][5]

随机森林是一个包含多个决策树的分类器。

为了提高随机森林的能力，分别采用信息增益、信息增益率、基尼系数作为划分特征的依据。信息增益相关计算公式分别为信息熵、条件熵，信息熵计算公式如下：

$$\mathrm{Info}(D) = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{5.5}$$

p_i表示选择该分类的概率，k表示分类的数目，本题$k = 2$。当引入另一个变量 A（特征 A）时，其水平变量 A 的各个水平所分割。通过计算在变量 A 的各个水平下随机变量 D 的信息熵的加权而得知在引入随机 A 后随机变量 D 的混乱程度，这一指标称条件熵，其计算公式如下：

$$\text{Info}(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}\times\text{Info}(D_i)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{i=1}^{k}\frac{|D_{ik}|}{|D|}\log_2\left(\frac{|D_{ik}|}{|D|}\right) \tag{5.6}$$

i表示某个引入变量 A 的水平，n表示变量 A 的水平个数，D 表示随机变量 D 的观察总数，D_i表示被随机变量 D 在变量 A 的i水平所分割的观测数。信息增益，其表达式如下：

$$Gain(D|A)=\text{Info}(D)-\text{Info}(D|A) \tag{5.7}$$

信息增益率计算公式如下：

$$\text{GainRate}(D|A)=\frac{\text{Gain}(D|A)}{\text{Info}(A)}=\frac{\text{Info}(D)-\text{Info}(D|A)}{\text{Info}(A)} \tag{5.8}$$

其中，$\text{Gain}(D|A)$表示在特征 A 条件下目标变量 D 的信息增益，$\text{Info}(A)$表示特征 A 的熵。

基尼系数计算公式如下：

$$\text{Gini}(p)=\sum_{i=1}^{k}p_i(1-p_i)=1-\sum_{i=1}^{k}{p_i}^2 \tag{5.9}$$

p_i表示选中的样本属于k类别的概率，则这个样本被分错的概率是$(1-p_i)$。对于一个样本集合 D，特征 A 对数据集合 D 的基尼系数为:

$$\text{Gini}_A(D)=\sum_{i=1}^{k}\frac{|D_i|}{|D|}\times\text{Gini}(D_i) \tag{5.10}$$

2. 重要性特征提取模型的建立与求解

读取降噪后的数据，在“主成分特征提取模型”的基础之上运用随机森林算法建立“重要性特征提取模型”。从“重要性特征提取模型”中得到 16 个特征，即得到维度为 100×16 的样本特征数据，每一行表示每个样本特征提取后的特征数据。其部分结果如表 4 所示，其完整结果保存在附件“附表 2-1.csv”的和“附表 2-2.csv”中。

表 4 样本特征数据

样本序号	指标 1	指标 2	…	指标 15	指标 16
1	1.00085	0.76056	…	-1.89916	-0.39348
…	…	…	…	…	…
50	0.40550	-0.83522	…	-0.19953	0.40105
…	…	…	…	…	…
100	0.16764	0.37143	…	0.41191	-2.36363

第三部分：问题 3 的 1 个模型

（一）K-均值聚类算法模型

1. K-均值聚类算法理论

先随机选取 K 个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。一旦全部对象都被分配了，每个聚类的聚类中心会根据聚类中现有的对象被重新计算。直到聚类中心不再发生变化。本题 K 为 2。

2. 模型的建立与求解

根据题意，我们先新添的 3 项指标分别为精确率、F1-score、错误率。读取问题 2 维度为 100×16 的样本特征数据。我们将构建的模型训练并测试 100 次。每一次构建模型，将其中一个未测试过的样本特征数据作为测试集，将剩下 99 个样本中 70%作为训练集，30%作为验证集。实验结果表明，序号为 1、4、7、10、12、14、17、20、21、23、25、27、28、30、31、32、35、37、40、45、47、50、52、55、56、57、59、61、62、68、69、74、75、79、81、82、86、87、90、94、97、100 的 42 个样本在该模型下分类错误。最终得到改模型分类结果 6 个指标，其部分结果如表 5 所示，其完整数据保存在附件“K-均值聚类算法结果.csv”中。

表 5 K-均值聚类算法结果

序号	准确率/%	召回率/%	耗时/ms	精确率/%	F1-score	错误率/%
1	0	0	744.257	0	0	100
2	100	100	701.240	100	1.0	0
…	…	…	…	…	…	…
100	0	0	794.245	0	0	100
Mean(均值)	58%	58%	722.030ms	58%	0.58	42%
Std(标准差)	0.494	0.494	89.54	0.494	0.494	0.494

由表 5 可知，K-均值聚类算法分类预测结果远达不到题意要求，不能用该算法做分类预测。因此，采用混合高斯模型进行建模。

（二）混合高斯模型

1. 高斯模型理论[5][6]

单高斯模型：当样本数据 X 是多维数据时，高斯分布遵循概率密度函数：

$$P(x|\theta)=\frac{1}{(2\pi)^{\frac{D}{2}}}\exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right) \tag{5.11}$$

其中,μ为数据均值（期望），Σ为协方差，D 为数据维度。高斯混合模型可以看作是由 K 个单高斯模型组合而成的模型，K 个子模型是混合模型的隐变量。一个混合模型可以使用任何概率分布，使用高斯混合模型是因为高斯分布具备很好的数学性质以及良好的计算性能。

2. 模型的建立与求解

根据题意，我们先新添的 3 项指标分别为精确率、F1-score、错误率。读取问题 2 维度为 100×16 的样本特征数据。我们将构建的模型训练并测试 100 次。每一次构建模型，将其中一个未测试过的样本特征数据作为测试集，将剩下 99 个样本中 70%作为训练集，30%作为验证集。实验结果表明，序号为 1、10、20、21、32、47、56、68、79 的 9 个样本在该模型下分类错误。最终得到改模型分类结果 6 个指标，其部分结果如表 6 所示，其完整数据保存在附件“附表 3.csv”中。

表 6 混合高斯模型分类结果

序号	准确率/%	召回率/%	耗时/ms	精确率/%	F1-score	错误率/%
1	0	0	359.822	0	0	100
2	100	100	399.499	100	1.0	0
…	…	…	…	…	…	…

100	100	100	383.208	100	100	100
Mean(均值)	91%	91%	359.573ms	91%	0.91	9%
Std(标准差)	0.29	0.29	26.97	0.29	0.29	0.2862

该模型分类预测准确率为 91%，准确率标准差为 0.29，满足题意要求。

（三）问题 3 的两种数学模型的比较

“K-均值聚类算法模型”虽然简单且运算较快，但是其容易受到“孤立点”的影响。虽然经特征提取后得到 16 个特征数据，但是在该问题上分类准确率较低。因此，K-均值算法在该问题上效果不佳。“混合高斯模型”用高斯概率密度函数精确地量化事物，将一个事物分解为若干的基于高斯概率密度函数，混合模型不仅可以解决高纬度数据问题，还能够使用任何概率分布，使得该模型的能力增强。其运算速度快，准确率也高。

第四部分：问题 4 的 2 个模型

（一）基础分类器

1. 常见分类器

本模型所用分类器包括：逻辑回归、决策树、梯度提升决策树、AdaBoost、线性判别分析、二次判别分析、支持向量机、XGBoost、Voting Classifier、随机森林。

我们新添 3 项指标分别为精确率、F1-score、错误率。读取问题 2 维度为 100×16 的样本特征数据，为样本特征数据的两类故障分别加上标签。利用上述分类器，对样本特征数据做二分类。从 100 个样本特征数据随机选取 70%作为该模型的训练集，其余的 30%作为测试集来评判模型的分类准确率，其分类结果如表 7 所示，结果保存在“十个分类器分类结果.csv”。

表 7 十个分类器分类结果

分类器	准确率/%	精确率/%	召回率/%	F1-score
逻辑回归	53	49	53	51
随机森林	90	90	91	90
决策树	63	65	63	64
梯度提升决策树	87	87	86	87
AdaBoost	87	86	87	86
线性判别分析	53	49	53	51
二次判别分析	83	86	83	80
支持向量机	87	86	87	87
XGBoost	87	87	87	86
Voting Classifier	83	80	83	83
高斯朴素贝叶斯分类器	90	91	90	90

根据实验结果，依据准确率、精确率、召回率以及 F1-score 综合评判，我们从中选取随机森林分类器和高斯朴素贝叶斯分类器。

（二）随机森林模型

1. 模型的建立与求解

在无优化以及无调参的情况下，随机森林分类器[7]未能满足题意要求，因此我们运用控制变量法对模型参数进行调整，进而通过模型得分准则得出最佳分类器。首先我们对参数 n_estimators 进行调整，利用控制变量法，在模型得分中可得模型得分最高的参数值。

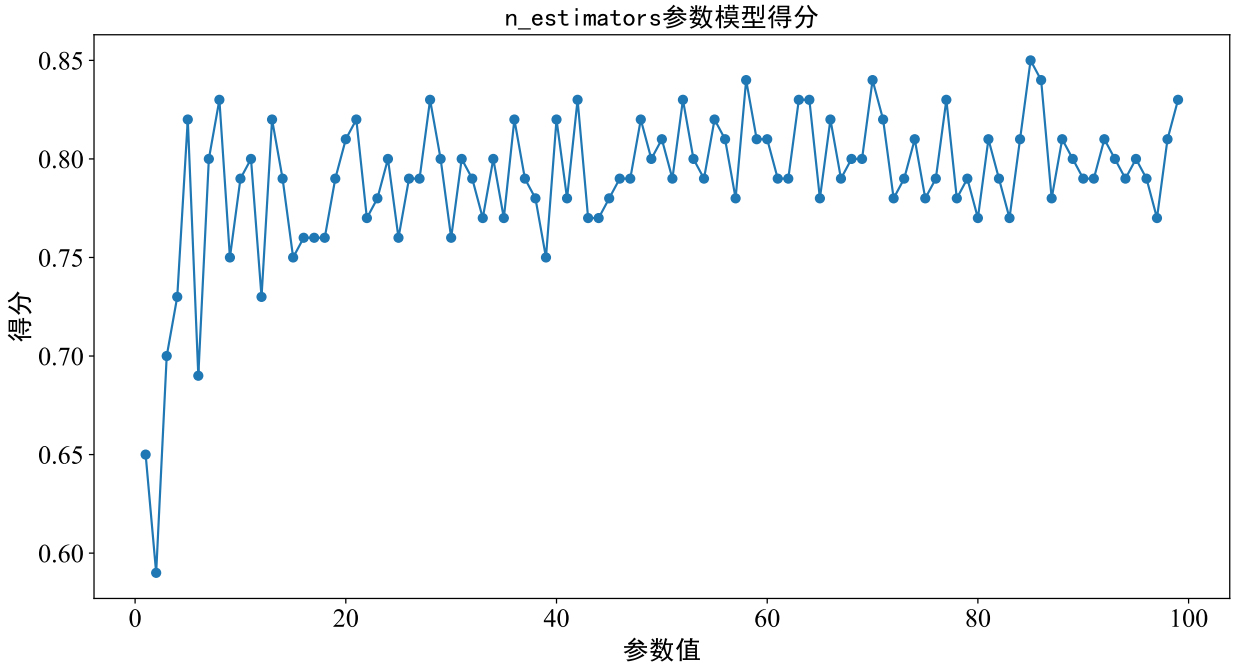


图 5-5 n_estimators 参数调整

如图 5-5 所示。通过模型得分可得当参数 n_estimators 为 85 时，得分最高且为 0.85。

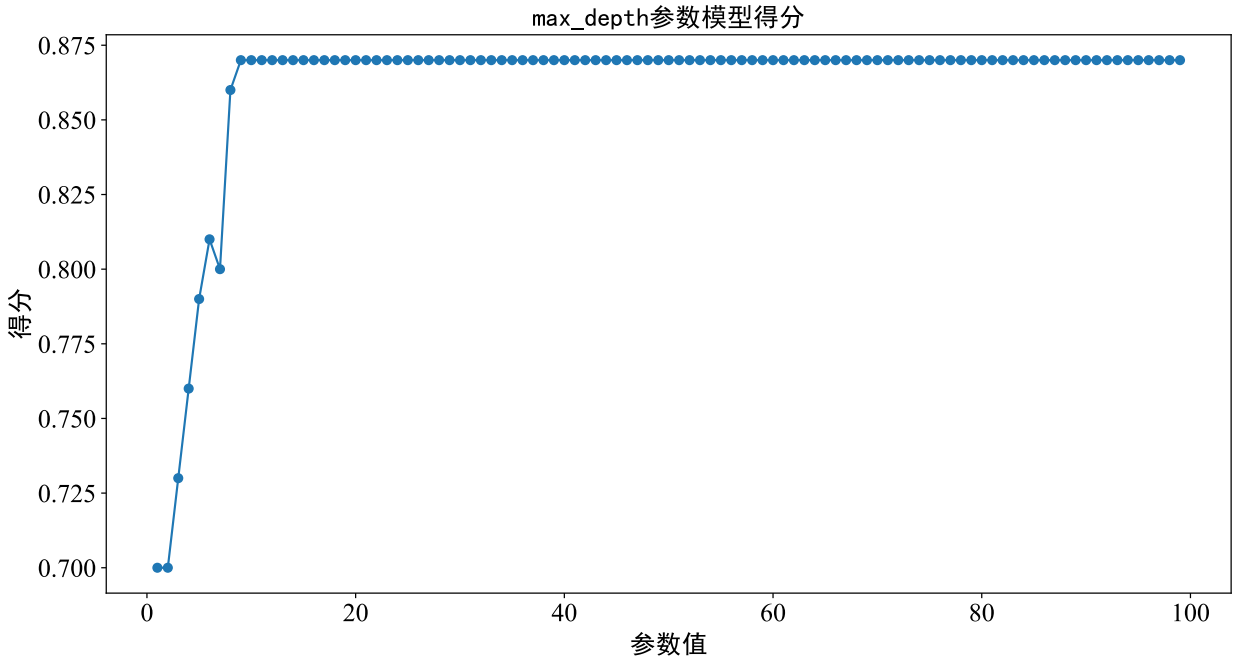


图 5-6 max_depth 参数调整

如图 5-6 所示，当 n_estimators 为 85 且其余参数为默认时，只改变 max_depth 的值，可得当参数 max_depth 为 20 时得分最高且为 0.87。

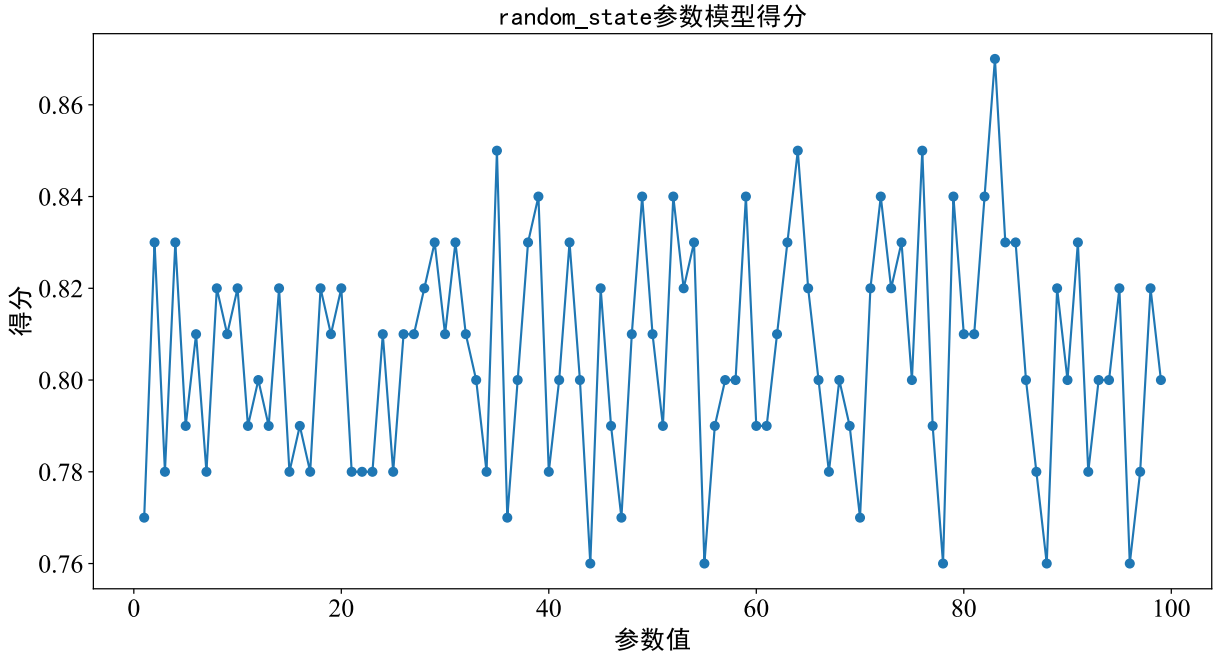


图 5-7 random_state 参数调整

如图 5-7 所示，当 n_estimators 为 85、max_depth 为 20 且其余参数为默认时，只改变 random_state 的值，可得当 random_state 为 83 时得分最高且为 0.87.52。

利用调整好的参数得到最佳随机森林分类器，将该分类器对样本特征数据进行分类，得到分类结果准确率为 93.33%，对比未优化的随机森林分类器，最佳随机森林分类器分类准确率有所提升。

我们用此分类器对 100 个样本特征数据进行单一样本分类，且每个样本预测 100 次。在预测过程中，我们将其中一个未测试过的样本特征数据作为测试集，将剩下 99 个样本中 70%作为训练集，30%作为验证集。最终所得部分结果如表 8 所示。其完整结果保存在附件“随机森林附表.csv”的附表中。由该附表可知，编号为 10、50、71、76、79、80、86、90、94、100 的样本分类结果错误，并且预测准确率的均值为 90%，未能达到题意要求。

表 8 随机森林模型分类结果

序号	准确率/%	召回率/%	耗时/ms	精确率/%	F1-score	错误率/%
1	100	100	1558.75	100	1.0	0
2	100	100	1548.68	100	1.0	0
…	…	…	…	…	…	…
100	0	0		0	0	100
Mean(均值)	90%	90%	1673.70ms	90%	0.90	10%
Std(标准差)	0.30	0.30	114.20	0.30	0.30	0.003

（三）高斯朴素贝叶斯模型

1. 高斯朴素贝叶斯分类器理论[8][9]

高斯分布概率密度函数为(5.11)，朴素贝叶斯公式如下：

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{\sum_1^n P(x|y_i)P(y_i)} \tag{5.12}$$

2. 模型的建立与求解

虽然经过优化后得到最佳随机森林分类器，但其预测准确率均值达不到题意要求。因此，我们需要寻找一个新的分类器进行建模。在最佳随机森林分类器的分类准确率达到 93.33%的基础上，我们可以得出的结论是：将原始样本数据经过特征提取后得到的样本特征数据没有问题。因此，我们在更改分类器的同时，继续用样本特征数据进行分类。最佳随机森林分类器也许对我们所提取的样本特征数据并没有其他分类器敏感。因此，我们根据十个分类器的分类结果，选择和随机森林分类准确率相同的高斯朴素贝叶斯分类器。

我们构建高斯朴素贝叶斯分类器时进行微小的调整，测试集我们选取所有数据的 20%，最后得出的分类准确率为 95%。在此基础上，我们用此分类器对 100 个样本特征数据进行分类，且每个样本都预测 100 次。在预测过程中，我们将其中一个未测试过的样本特征数据作为测试集，将剩下 99 个样本中 80%作为训练集，20%作为验证集。最终所得部分结果如表 9 所示。其完整结果保存在附件“附表 4.csv”中。可以看出其中编号为 10、79、100 的分类结果是错误的，并且预测准确率的均值为 97%，符合题意要求。

表 9 高斯朴素贝叶斯模型分类结果

序号	准确率/%	召回率/%	耗时/ms	精确率/%	F1-score	错误率/%
1	100	100	277.675	100	1.0	0
2	100	100	360.433	100	1.0	0
…	…	…	…	…	…	…
100	0	0	257.178	0	0	100
Mean(均值)	97%	97%	311.337ms	97%	0.97	3%
Std(标准差)	0.17	0.17	56.44	0.17	0.17	0.0017

（四）问题 4 的两种数学模型的比较

1. 随机森林模型优点：
   1) 随机森林可以处理高维度的数据；
   2) 创建随机森林时，对 generalization error 使用无偏估计，增强模型泛化能力；
   3) 对于不平衡数据集，随机森林可以平衡误差。当存在分类不平衡的情况时，随机森林能提供平衡数据集误差的有效方法；
   4) 虽然特征提取过程会有遗失部分特征，但是随机森林仍然可以维持较高准确率；
   5) 随机森林算法有很强的抗干扰能力和抗过拟合能力。
2. 随机森林的缺点：
   1) 随机森林就像一个黑盒子，无法控制模型内部的运行。只能在不同的参数之间进行调试而不能修改核心算法。
   2) 100 个样本数据属于小样本，数据量不足也影响最终分类结果。
   3) 相对本文其他分类器，该分类器运算速度较慢。对单一样本预测 100 次所用的时间在 1700ms 左右。
3. 高斯朴素贝叶斯分类器模型优点：

1) 朴素贝叶斯模型源于古典数学理论，有稳定的分类效率。
2) 对小规模的数据表现很好，能处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。
3) 该分类器在二分类数据上表现优异。对 100 个样本测试集 30%，训练集 70%，得出的准确率为 95%，最后的预测准确率均值高达 97%。
4) 该分类器的运算速度较快。对单一样本预测 100 次所用的时间在 270ms 左右。

4. 高斯朴素贝叶斯分类器模型缺点：

1) 需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会受假设的先验模型的影响而导致预测效果不佳。

第五部分：特征数量和质量对特征识别的影响以及解决方法

（一）特征数量以及质量对特征识别的影响

在贡献率达到 95%以上，选出的特征共 72 个，用 100 个样本把全部的 72 个特征都囊括，测试集样本和训练集样本分别占 30%和 70%，。随后用随机森林模型和高斯朴素贝叶斯模型分类得出准确率为 83.33%和 86.67%。我们选取随机森林分类器来判断特征的质量，在随机森林分类器中做单一属性分类，分类准确率部分结果如表 10 所示，其完整结果保存在附件“RF_72 单一属性分类.csv”中。

表 10 单一属性分类准确率

特征序号	精确率/%	召回率/%	F1-score	准确率/%
18	89	87	85	87
29	79	80	78	80
…	…	…	…	…
67	58	57	57	57
…	…	…	…	…
41	45	43	44	43

（二）解决方法

从表 10 可以看出每一个特征对于分类而言质量的优劣，为了提高最后的分类准确率，我们从 72 个特征中进行特征选取，利用随机森林分类器，将准确率从高到底依次排序组合，组合特征分类准确率如表 11 所示。最后选取前 16 列特征得到分类的最佳效果。

表 11 组合特征分类准确率

前 n 个特征组合	精确率/%	召回率/%	F1-score/%	准确率/%
1	89	87	85	87
2	84	83	84	83
…	…	…	…	…
10	83	83	82	83
…	…	…	…	…
16	91	91	90	90
…	…	…	…	…
72	89	87	85	87

六、结果分析

总体来说，本文实验结果真实可靠，每个问题都是紧密相连。其正确性和合理性在每个问题中都得到了验证。

1. 问题 1 构建的“箱线图异常点调整模型”合理运用箱线图挖掘异常点的优势，并根据实际情况进行调整，即保留原始数据，又达到降噪效果，如图 6-1 部分降噪后数据所示，是图 5-2 降噪后的结果。

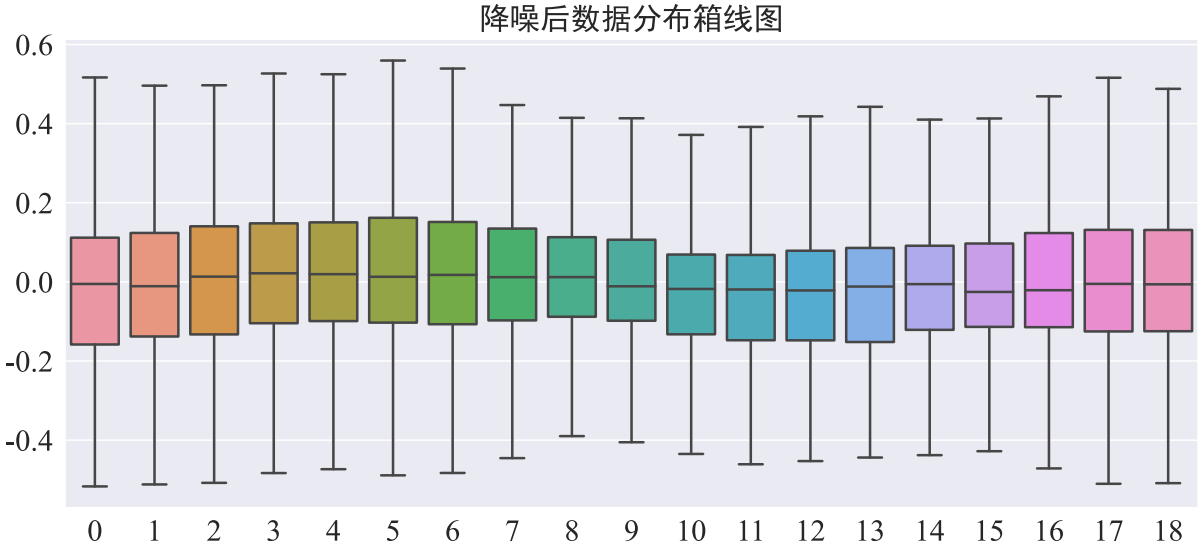


图 6-1 部分降噪后数据

2. 构建的“主成分特征提取模型”提取出 72 个重要特征。在此基础上，再构建“重要特征提取模型”，提取出 16 个更为重要的特征。

3. 问题 3 运用问题 2 提取的样本特征数据进行分类，经过两次建模，预测准确率均值最终可达 91%。不仅证明箱线图降噪和特征提取方法的正确性和合理性，同时也证明了“混合高斯模型”的正确性和合理性。

4. 问题 4 同理运用问题 2 提取的样本特征数据进行分类，经过两次建模，预测准确率均值最终达到 97%。再次证明上述方法和“高斯朴素贝叶斯模型”的正确性和合理性。

5. 问题 5，我们经过重复构建上述模型进行分析和比较，得出的结果一致。

6. 同时，我们发现序号为 1、10、20、21、32、47、56、68、79 的 9 个样本在“混合高斯模型”中分类错误；序号为 50、71、76、78、79、80、86、90、94、100 的 10 个样本在“随机森林模型”中分类错误；序号为 10、79、100 的 3 个样本在“高斯朴素贝叶斯模型”中分类错误。无论是哪个模型，10、79 号样本分类都是错误。由此，可推测数据可能存在异常。

七、模型评价与推广

（一）模型的优点

我们最终的模型都得到了满意的结果。

1) 从“箱线图异常点调整模型”中，我们把噪声数据进行适当调整，解决了噪声对分类的影响。

2) 从“主成分特征提取模型”中，我们从 4096 列数据中得到数据的 72 个主成分

特征，解决了从原始数据中对特征的提取问题。

3) 从“重要特征提取模型”中，我们从 72 个主成分特征提取分类效果最佳的 16 个特征，既降低数据维度、减少计算量，又解决了预测准确率均值不达标的问题。
4) 从“混合高斯模型”中，我们得到最终的预测准确率均值达到 91%，对比“K-均值聚类算法模型”67%的准确率，解决了准确率不高的问题。
5) 从“高斯朴素贝叶斯模型”中，我们得到最终的预测准确率均值达到 97%，解决了“随机森林模型”预测准确率均值不达标的问题。同时减少了计算量，从原来的对一个样本的 100 次分类 1700ms 左右降低到 200ms 左右。

（二）模型的缺点

1) 在对特征进行优劣评判的时，只用了随机森林去评判特征的优劣，没有综合其他分类器的评判。
2) 在对特征的综合选择上我们只将准确率从高到低依次排序组合，选取特征的唯一标准就是准确率，没有用随机组合来评判组合特征的优劣。因此，在特征选择上可能存在一定的偏差。
3) 由于我们构建模型时未对算法本身进行大幅度改进，所以也许分类准确率还未达到最高。

（三）模型的改进

1) 在对特征优劣的评判时，我们可以用多种适合此类数据的分类器去分类，然后综合各个分类器的分类结果对特征进行优劣排序。
2) 在对特征进行组合时，我们可以把 72 个特征每一种组合的方式都用多种分类器去进行分类，最后确定是哪一种组合最佳。但是由于工作量太大未能实现。
3) 由于原始数据是信号数据，可以在数据预处理阶段使用非传统的滑动窗口提取特征数据。
4) 在提取特征时，我们可以把第一类 70 个样本数据使用皮尔逊不相关性系数，得到 70 行数据的皮尔逊不相关性距离矩阵。再利用 Python 中的 Giotto-TDA 拓扑机器学习工具箱对数据进行网络构建和 VR 复形滤流的构建。对于 VR 滤流得到的持续图 PD 并不能直接进入机器学习算法进行分类，所以基于拓扑数据分析的现有方法使用持续同调方法。将持续图 PD 和条形码转换为持续熵，持续景观图 PL、贝蒂曲线 BE、热核图 HE，这三种图能从不同的维度观察持续图 PD 的特点。相对于直接从复杂的信号数据中提取特征进行分类，从持续图 PD 中提取拓扑特征具有极高的可解释性与可视化性。为了高度概括三种拓扑特征图，我们使用矩阵 1-范数和 2-范数对 PL,BE,HE 以及持续熵进行总结，在 VR 复形的每一维获取一个矩阵范数，从而获得可用于机器学习分类的特征数值。
5) 对于分类算法，我们之前利用了滑动窗口增加了样本数据，可能会在一定程度上增加训练模型的分类准确率。但是随着样本的增加，可让所有的滑动窗口和原始每一个样本数据联系起来，提高分类准确率。

八、参考文献

[1] 顾国庆, 李晓辉. 基于箱线图异常检测的指数加权平滑预测模型[J]. 计算机与现代化, 2021, 1: 28-33.
[2] 周志华. 机器学习. 北京：清华大学出版社，2016 年，229-232
[3] Abdi H, Williams L J. Principal component analysis[J]. Wiley interdisciplinary reviews: computational statistics, 2010, 2(4): 433-459.
[4] Criminisi, A., Shotton, J., & Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and trends® in computer graphics and vision, 7(2–3), 81-227.
[5] McLachlan G J, Rathnayake S. On the number of components in a Gaussian mixture model[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2014, 4(5): 341-355.
[6] Reynolds D A. Gaussian mixture models[J]. Encyclopedia of biometrics, 2009, 741(659-663).
[7] Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions[J]. ISPRS journal of photogrammetry and remote sensing, 2016, 114: 24-31.
[8] Jahromi A H, Taheri M. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features[C]//2017 Artificial Intelligence and Signal Processing Conference (AISP). IEEE, 2017: 209-212.
[9] Cataldi L, Tiberi L, Costa G. Estimation of MCS intensity for Italy from high quality accelerometric data, using GMICEs and Gaussian Naïve Bayes Classifiers[J]. Bulletin of Earthquake Engineering, 2021, 19(6): 2325-2342.

九、附录

1. 由于我们是在 PyCharm 搭建 Python3.8 编译环境下运行代码，其他环境下结果会存在一定的差异。
2. 本文所用 Python 主要库包含以下：Numpy19.2、Pandas1.3.3、Matplotlib3.4.3、seaborn0.11.2、scikit-learn1.0.2、generator1.147。
3. 解决问题的所有源代码、附表、图都在附录压缩包内。
4. 各附表说明如表 12 所示。

表 12 附表说明

文件名	说明
100.csv	从 txt 读取并整理的数据
A_denoising_data.csv	100.csv 降噪后的数据
附表 1-1.csv	问题 1 的结果
附表 1-2.csv	问题 1 的结果
contribution_rate_100.csv	主成分分析贡献率排序
contribution_rate_100_sum.csv	主成分分析前 n 个特征贡献率累计
AandB_feature_NO_lable_data.csv	72 个特征数据
十个分类器分类结果.csv	十个分类器分类结果
RF_72 单一属性分类.csv	随机森林单一属性分类结果
附表 2-1.csv	问题 2 的结果
附表 2-2.csv	问题 3 的结果
K-均值聚类算法结果.csv	K-均值聚类算法结果
附表 3.csv	问题 3 的结果
随机森林附表.csv	随机森林分类结果
附表 4.csv	问题 4 的结果