所在组别	2022 年第二届中国高校大数据挑战赛	参赛编号
本科组		bdc220279

基于决策树及集成方法的机械设备故障检测与规则挖掘

摘要

随着国家科学技术的提升，工业的发展也是日新月异。伴随着每次工业的巨大进步，我们的工业制造核心机械设备也是向着自动化和精密化发展，在机械设备的服务周期内，难免不会出现老化甚至是无法工作等问题，严重影响了生产的效率，增大了生产的成本。本文运用决策树模型和集成算法对机械设备的故障进行了预测，并且分析了故障设备的具体故障原因，挖掘了其中潜在的规则。

针对问题 1: 我们首先对数据进行了可视化，利用题目条件进行了**异常值剔除**，利用数据分布进行了**偏度变换**，利用相关系数删除了无关变量“机器编号”和“统一规范编码”。对数值不敏感变量“室内温度”、“设备温度”进行了**最大最小归一化**，对分类型变量“机器质量等级”进行了**独热编码**。其次我们进行了特征工程，根据“转速”和“扭矩”之间明显的函数关系，我们构造了两个新特征：“转速扭矩乘积”和“相对残差”，利用“室内温度”和“设备温度”构造了新特征“温差”，观察正常、异常样本“使用时长”的频率分布直方图，构造了“时长 >200min”。最后我们利用 **XGBoost** 进行训练，利用树模型特有的**信息增益**找出了最重要的五个指标：**“转速”**、**“扭矩”**、**“使用时长”**、**“温差”**、**“转速扭矩乘积”**。

针对问题 2: 我们首先分析评判指标，在实际生产过程中，机械设备故障往往是较为严重的问题，需要及时解决，因此在进行预测时应尽可能将故障样本找全，追求高**召回率**；此外，如果近为了找全样本而不顾查准率，则会增加成本，因此我们也要追求高**精确率**。为了平衡两个指标，我们选用二者的调和平均数$\boldsymbol{F1-score}$ 作为评判指标。对于预测任务，我们先使用网格搜索法寻找最优参数，训练了五个基学习器：**XGBoost**、**随机森林**、**CatBoost**、**LightGBM**、**ExtraTree**，然后使用使用集成算法 **Stacking**、**Voting** 分别构建模型，最后比较这些模型的性能，Voting 性能最佳，在 5 折交叉验证下的 $F1$ 为 **0.8123**。

针对问题 3: 本问为多分类问题，由于存在样本不平衡问题，我们选用适用于多分类的 $macro-F1$ 作为评价指标。由于异常样本数较少，我们选用易于处理小样本且易于解释的模型**决策树**作为预测模型，使用网格搜索法寻找最优参数，得到 5 折交叉验证下的 $macro-F1$ 为 **0.8337**。

针对问题 4: 本问要求对所给数据进行预测，我们采用问题 2 和问题 3 中相同的数据处理方法，利用已经求得的最优模型进行训练，对于附件中的 1000 条数据，我们一共找出 **30** 个异常样本，并对故障原因进行了预测，**图 13**列出了其中 10 个。

针对问题 5: 我们利用问题 3 中已经训练好的决策树进行可视化，生成的树如**图 12**所示，我们认为使得某一类成为主要类的第一个分叉节点是主要影响节点，得到 HDF、OSF、PWF、RNF、TWF 的主要特征属性为 **“温差”**、**“扭矩”**、**“温差”**、**“转速扭矩乘积”**、**“扭矩”**，并且分析了主要成因，挖掘了潜在的规则如**表 4**所示。

关键词: 特征工程　信息增益　集成学习　决策树　样本不平衡　规则挖掘

1 问题重述

1.1 问题背景

随着我国科学技术和生产力的高速发展，我国工业化发展也上了快车道，是我国经济乃至世界经济的支柱。然而伴随着工业规模的逐渐扩大，工厂中使用的机械设备也逐渐向着两个方向发展，巨型化和精密化，但不论是那种方向，机械设备故障的检测难度是越来越大，传统的人工检测或许不能分析到全面的信息。但机械设备出故障后带来的后果往往是我们无法承担或是不愿承担的，随着工业化的扩大，机械设备的负荷量越来越大，零件和耦合部分越来越复杂，往往一个零件的破损就会使得整个设备崩溃，对企业和社会都会带来不可估摸的影响。因此，充分使用现代技术来实现预测工艺的科学化，自动化，智能化是十分重要且必要的。准确高效地预测工业机械设备中的故障，是新时代工业的要求，也是我们进一步发展工业不可忽视的一步。

1.2 问题要求

本题要求我们对工业机械设备故障进行预测，数据提供了 9000 例样本，其中工业机械设备常见的指标如室温，机器温度，转速，扭矩和使用时长等等，而机器故障的类型又具体分为 5 种。我们需要建立模型分析是否故障以及故障类型，然后根据我们模型来探究各种故障类型与特征的关系，挖掘特征与故障之间的联系。

问题 1：进行数据预处理，分析数据并选择合适的预测指标。
问题 2：建立模型与评价方式，预测机械设备是否发生故障。
问题 3：建立模型与评价方式，预测机械设备发生故障的具体故障类型。
问题 4：利用问题 2、3 的模型对所给数据进行预测。
问题 5：分析各故障类型的主要成因，挖掘潜在模式或规律。

2 模型假设

- 假设各样本之间独立同分布。
- 假设不同样本采用相同的测量手段采集信息。
- 假设某些变量间的函数关系可用初等函数表示。

3 符号说明

符号	说明
g_i	损失函数对预测值的一阶导数
h_i	损失函数对预测值的二阶导数
λ	正则化系数
a_*	当前最优划分特征
a_*^v	最优特征中的某一取值
$\mathfrak{L}$	基学习器算法
h	基学习器输出
H	元学习器输出
n	样本数量
x_i	特征 i
y	标签
$\hat{y}$	预测值

4 模型建立和求解

4.1 问题一模型建立与求解

4.1.1 问题分析

数据集中有“机器编号”, “统一规范代码”, “机器质量等级”, “室温”, “机器温度”, “转速”, “扭矩”, “使用时长”, “是否发生故障”, ” 具体故障类别” 共 9 列数据，其中前七个是样本特征，后两个是标签。出于在现实生活中的理解，机器编号和统一规范代码应该对于分类任务来说是无效特征，因此数据预处理是必不可少的。在对原本的特征进行筛选和融合后，得到新的特征作为我们后续解决问题的基础。

4.1.2 数据预处理与特征工程

首先对数据进行初步的探索，对数据缺失值进行统计，发现所有特征均没有缺失值。然后对哑变量“机器质量等级”进行转换，去除“统一规范编码”的首字母，完成初步处理。

注意到标记为 1 的样本均为异常样本，相应的异常类型不会是 “Normal”，因此剔除下列两个异常值:

	机器编号	统一规范代码	机器质量等级	室温 (K)	机器温度 (K)	转速 (rpm)	扭矩 (Nm)	使用时长 (min)	是否发生故障	具体故障类别
7506	7729	L55686	L	298.4	309.6	1710	27.3	163	1	Normal
8015	3433	L56195	L	297.2	308.1	1431	49.7	210	1	Normal

图 1: 异常值剔除

接下来对数据分布进行探索，将变量两两之间的分布展示如下：

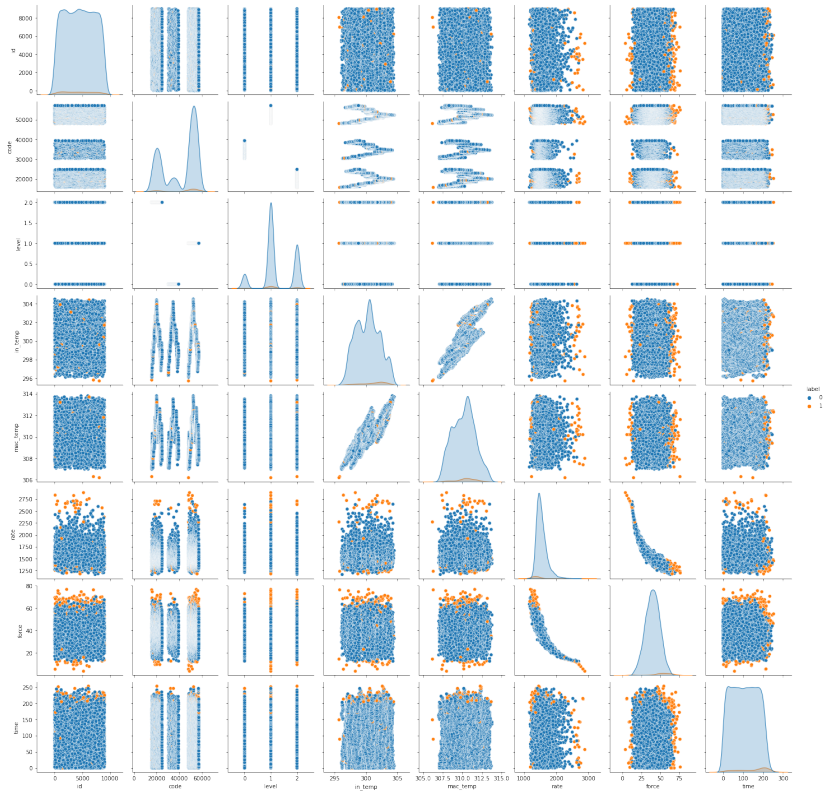


图 2: 变量两两间分布

变量之间的相关性展示如下：

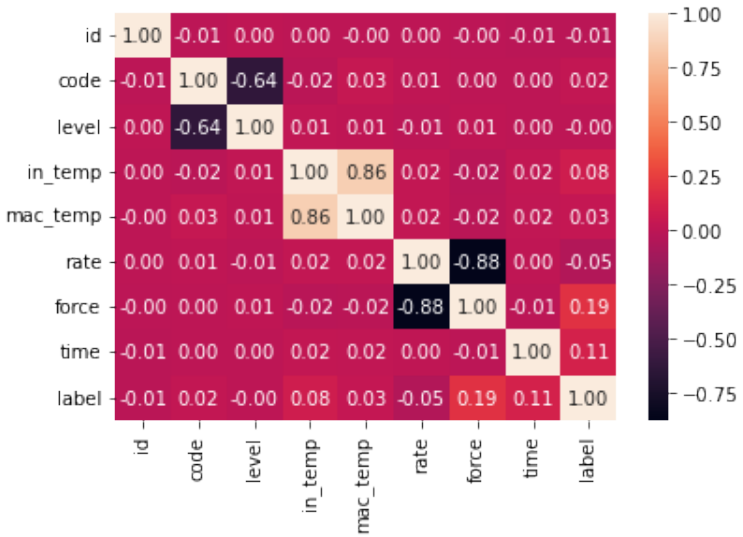


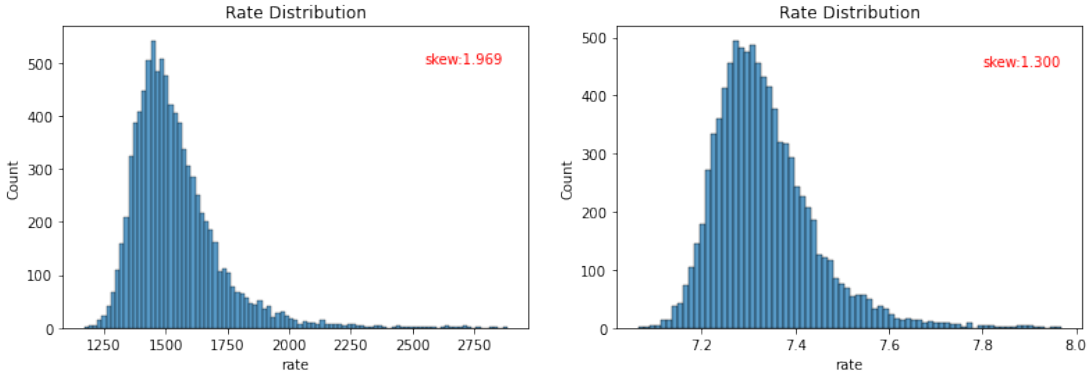
图 3: 变量相关系数热力图

从中我们可以发现，变量“机器编号”没有任何作用，变量“统一规范编码”与“机器质量等级”有关，本身不影响其他变量分布，故这两个变量可以删除。

注意到“室内温度”和“机器温度”这两个变量变化幅度很小，同时根据常识我们可以猜测机器温度与故障可能存在关系，因此我们增加一个特征“del_temp”表示机器温度与室内温度的差值，并且对“室内温度”和“机器温度”进行归一化处理，使得特征变化更明显。归一化公式如下：

$$x^{'} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

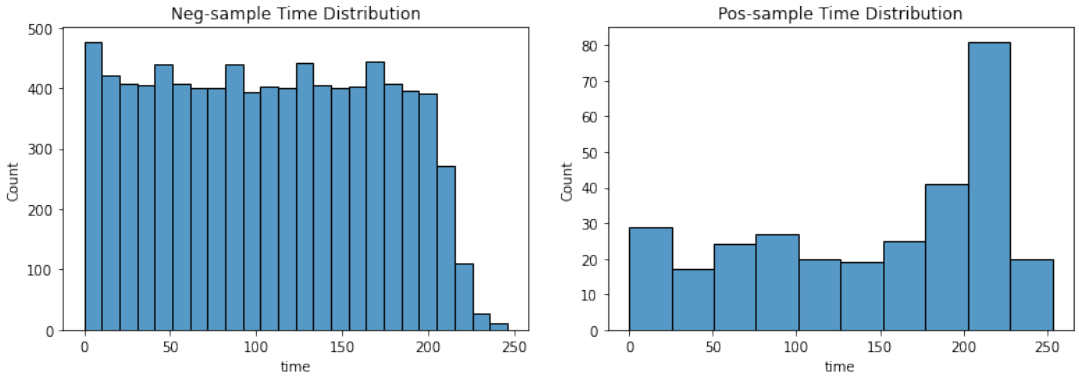
观察变量“转速”的分布，我们可以发现其存在明显的偏态分布，我们对其进行对数变换，使得分布更加趋向正态分布。



(a) 对数变换前分布　　(b) 对数变换后分布

图 4: 变量“转速”分布

观察变量“使用时长”的分布，可以发现负例样本在时长大于 200min 时数量急剧下降，正例样本在时长大于 200min 时数量急剧上升，因此我们增加特征“long_time”表示时长是否超过 200min。



(a) 正常样本使用时长分布　　(b) 故障样本使用时长分布

图 5: 使用时长分布

注意到变量“转速”和“扭矩”之间有明显的函数关系，趋势与反比例函数相似，因此我们增加二者的乘积作为新特征“multi”。同时趋势也与指数函数非常相似，且在正常样本中，同时随着“扭矩”的增大，数据分布逐渐变得分散，总体数据呈现“牛角状”。我们利用这个特性，首先使用正常样本进行回归，得到回归曲线，再求出所有样本与回归曲线的残差除以回归函数值作为新特征——

相对残差“res”，具体表达式如下：

$$\hat{y} = 7.386\times10^{8}e^{-2.286x} + 0.01142$$
$$res = \frac{y-\hat{y}}{\hat{y}} \tag{2}$$

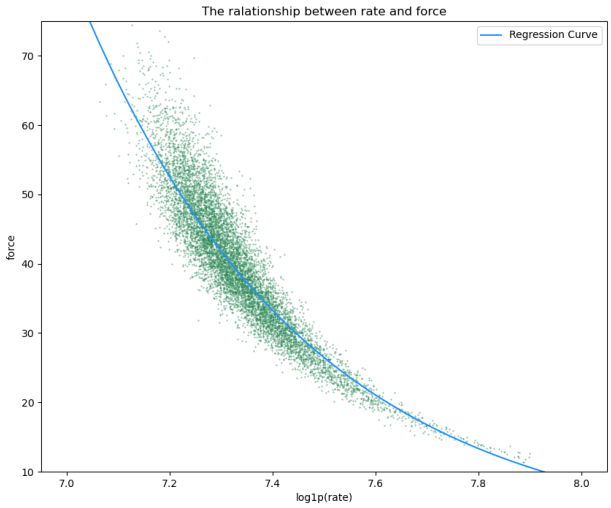


图 6: 转速、扭矩回归曲线

至此特征工程做完，处理后的部分数据展示如下：

	in_temp	mac_temp	rate	force	time	label	type	det_temp	multi	long_time	res	level_H	level_L	level_M
8160	0.204545	0.302632	7.421776	28.0	158	0	Normal	11.0	46788.0	0	-0.115152	1	0	0
3420	0.784091	0.526316	7.535830	21.3	87	0	Normal	7.6	39894.9	0	-0.126496	1	0	0
3758	0.886364	0.710526	7.290293	48.7	60	0	Normal	8.1	71345.5	0	0.139608	0	1	0
7538	0.318182	0.421053	7.345365	37.7	43	0	Normal	10.9	58359.6	0	0.000514	0	1	0
327	0.363636	0.513158	7.295735	42.6	164	0	Normal	11.2	62749.8	0	0.009339	0	1	0
3871	0.909091	0.815789	7.322510	40.1	135	0	Normal	8.7	60671.3	0	0.010055	0	1	0
1393	0.386364	0.276316	7.379008	31.4	121	0	Normal	9.2	50271.4	0	-0.100089	0	1	0
1168	0.420455	0.368421	7.570443	19.8	204	0	Normal	9.6	38392.2	1	-0.121195	0	0	1
2600	0.693182	0.578947	7.247081	46.8	151	0	Normal	8.8	65660.4	0	-0.007833	0	1	0
830	0.272727	0.184211	7.633854	19.2	150	0	Normal	9.5	39667.2	0	-0.014994	0	0	1

图 7: 处理后数据

4.1.3 XGBoost

为了筛选特征，我们必须使用一个量化指标来评判一个特征的重要性。在基于决策树的模型中，都具有评判特征重要性的指标：信息增益。因此，使用树模型进行训练，我们可以得到特征重要性的量化结果。在本任务中，我们使用 XGBoost 模型。

XGBoost 是一种集成学习算法，本质上是梯度提升树算法（GBDT）的改进，也是一种基于决策树的集成算法。

XGBoost 的目标函数如下：

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta) \tag{3}$$

其中 L 为损失函数，Ω 为结构风险，Θ 为模型参数。

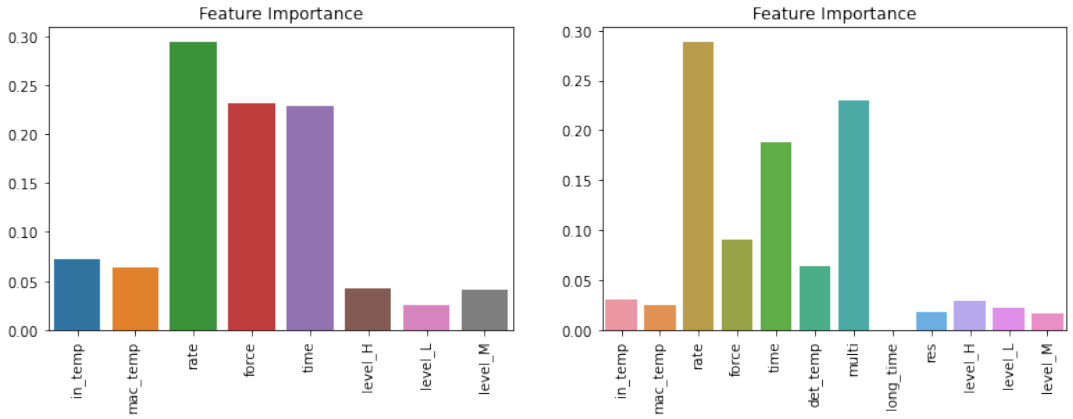
XGBoost 通过不断分裂生成树来构建一个集成模型，预测时将各个树的分数进行加权得到最后总分。XGBoost 中的核心算法，即分裂的过程如下：

Algorithm 1 Exact Greedy Algorithm for Split Finding [2]

Input: I,instance set of current node
Input: d,feature dimension
  $gain \leftarrow 0$
  $G \leftarrow \sum_{i \in I} g_i, H \leftarrow \sum_{i \in I} h_i$
  for $k = 1 : m$ **do**
    $G_L \leftarrow 0, H_L \leftarrow 0$
    for j *in sorted*$(I, by\ x_{jk})$ **do**
      $G_L \leftarrow G_L + g_j, H_L \leftarrow H_L + h_j$
      $G_R \leftarrow G - G_L, H_R \leftarrow H - H_L$
      $score \leftarrow \max(score, \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda})$
    end for
  end for
Output: Split with max score

4.1.4 问题一求解

我们使用 XGBoost 分别对特征工程前、特征工程后的数据进行训练，各特征信息增益如下：



(a) 特征工程前信息增益　　(b) 特征工程后信息增益

图 8: 信息增益

可以看到特征工程得到的变量“long_time”几乎无作用，主要特征为：转速、扭矩、使用时长、温差、转速扭矩乘积。

4.2 问题二模型建立与求解

4.2.1 问题分析

问题二中要求我们设计一个判别设备是否故障的模型，属于二分类问题。在问题一中我们进行了数据预处理和特征选择，并添加了新的特征。模型的输入应该有“室内温度”，“机器温度”，“转速”，“扭矩”，“使用时长”，“温差”，“使用时长是否超过 200min”，“转速 * 扭矩”，“残差”共九个特征，模型的输出我们定义 0 为正常，1 为异常。考虑到这是一个样本不均衡的分类任务，一般我们可以从三个层面来解决样本不均衡的问题

一是样本层面，这是一种从根本上解决样本不平衡的方法，包括了欠采样，过采样和组合采样等等。该方法的关键是如何去丢弃掉一些样本（或是如何根据已有的样本去添加样本），这种方法也容易造成模型的欠拟合 (或是过拟合)。

二是训练层面，我们给予少数类样本更大的权重，这样当模型在训练过程中就会更加重视少数类样本，从而减轻掉样本层面二者的不均衡，达到更好的训练效果。

三是模型层面，我们可以采用对全局信息不是很敏感的模型如决策树等模型，更好的选择是采用集成学习的方法。在每个基学习器中我们都使用降采样的方法使得两类样本均衡，每次训练基学习器时我们都会使错分的数据权值更大，到一定的训练次数后我们会得到一个分类性能好的强学习器。

在本问中，我们采用两种模型融合的集成方式：Stacking 和 Voting 求解。

4.2.2 Stacking

Stacking 是一种集成学习方法，能够将多个基学习器进行集成，以基学习器的输出作为元学习器的输入，最后输出结果，提高模型的精度。Stacking 算法如下：

Algorithm 2 Stacking 算法 [1]

Input: 训练集 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
初级学习算法 $\mathfrak{L}_1, \mathfrak{L}_2, \ldots, \mathfrak{L}_T$
次级学习算法 $\mathfrak{L}$.
for $t = 1, 2, \ldots, T$ **do**
  $h_t = \mathfrak{L}_k(D)$;
end for
$D' = \varnothing$;
for $i = 1, 2, \ldots, m$ **do**
  for $t = 1, 2, \ldots, T$ **do**
    $z_i t = h_t(x_i)$;
  end for
  $D' = D' \cup ((z_{i1}, z_{i2}, \ldots, z_{iT}), y_i)$;
end for
$h' = \mathfrak{L}(D')$;
Output: $H(x) = h'(h_1(x), h_2(x), \ldots, h_T(x))$

在本问中，我们使用随机森林、XGBoost、LightGBM、CatBoost 和 ExtraTree 五个基学习器，将他们的输出进行并联作为元学习器逻辑回归的输入，最后输出类别“是否故障”。模型结构如下：

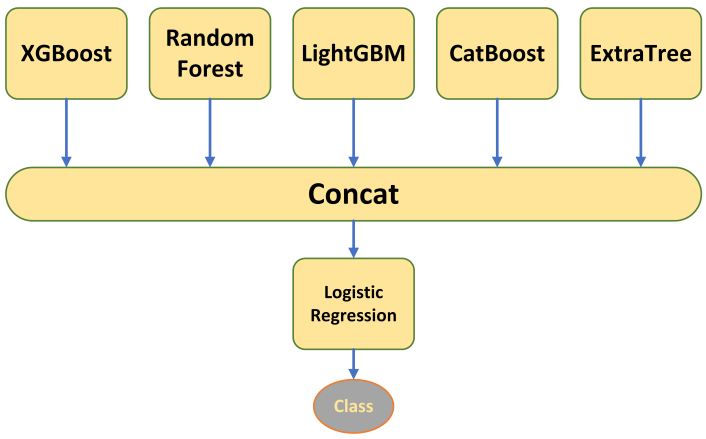


图 9: Stacking 模型结构

4.2.3 Voting

Voting 的思想非常简单，即让各基学习器投票进行预测。Voting 分为“硬投票”和“软投票”两种，前者把得票数最多的作为预测类别，而后者使用基学习器预测概率的均值，均值最大的类作为预测类别。

4.2.4 评价指标

在实际检修过程中，我们需要准确地确定机械设备是否发生故障，并且针对少数故障设备，我们应尽力找出，以免造成更大的后续影响。在概率统计中，我们有如下的混淆矩阵：

Reference	Prediction	
	$Positive$	$Negtive$
$Positive$	TP	FN
$Negtive$	FP	TN

表 1: 混淆矩阵

根据我们的任务要求，我们既要追求高“查准率”，又要追求高“查全率”，且二者重要性相当。查准率和查全率分别用 P 和 R 表示，$F1-score$ 表示二者的调和平均数，有如下表达式：

$$\begin{aligned} P &= \frac{TP}{TP+FP} \\ R &= \frac{TP}{TP+FN} \\ F1 &= \frac{2}{\frac{1}{P}+\frac{1}{R}} \end{aligned} \tag{4}$$

我们用 $F1$ 指标来作为模型的评判指标，当模型的 $F1$ 较高时，我们既能很好地找出故障设备，又能够找的比较准，误判概率低。

4.2.5 问题二求解

观察样本，我们发现正例与负例的比例为 301：8697，这说明样本非常不平衡，为此，我们在训练模型时将调大正例的权重，以便更好地找出正例。我们首先训练 RandomForest、XGBoost、LightGBM、CatBoost 和 ExtraTree，使用网格搜索得到 $F1$ 及最优参数如下所示：

模型	5 折交叉验证 $F1$	最优参数
XGboost	0.8112	'learning_rate': 0.3793, 'n_estimators': 500
RandomForest	0.5696	'max_depth': 6, 'min_samples_leaf': 5, 'min_samples_split': 2, 'n_estimators': 900
CatBoost	0.7539	'iterations': 1200, 'learning_rate': 0.0336
LightGBM	0.7786	'colsample_bytree': 0.75, 'learning_rate': 0.2336, 'n_estimators': 700, 'num_leaves': 16
ExtraTree	0.7285	'max_features': 5, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 1100

表 2: 五个模型最优参数

可以看到 XGBoost 表现最好，而随机森林表现远远不如其他模型，因此，我们进行模型融合时不再使用随机森林。我们利用训练好的最优模型分别进行 Stacking 和 Voting 分类器的构建，比较各模型 $F1$ 值，结果如下：

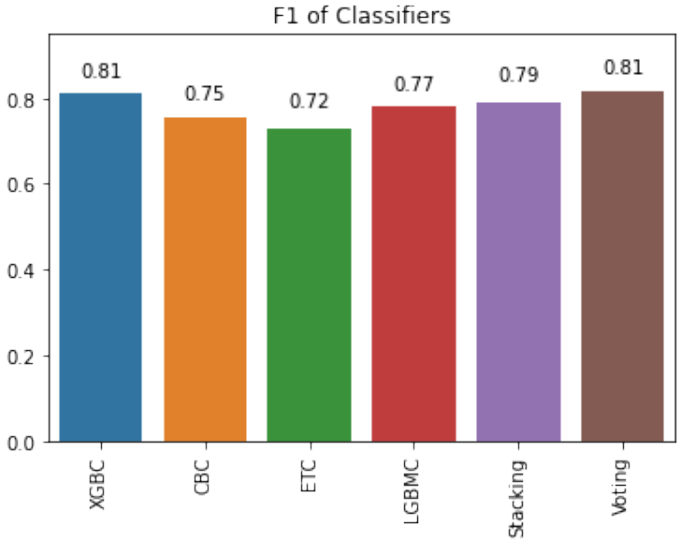


图 10: 各模型 $F1$ 比较

可以看到 Voting 表现最好，5 折交叉验证 $F1$ 值达到了 0.8123，Stacking 则为 0.7910，稍逊于基学习器中表现最好的 XGBoost。

4.3 问题三模型建立与求解

4.3.1 问题分析

本题需要我们设计模型用于判别机械设备发生故障的具体类别，是一个多分类任务。由于正常和故障两类样本的数量不平衡，加上我们在第二问中已经将正常类和故障类分开，故本题就剔除掉正常类的样本，只在故障样本中进行训练。同时我们注意到了问题五中需要分析各类故障的相关特征，并且要量化表示，因此我们希望多分类模型的决策过程逻辑是清晰的，故采用了决策树作为第三问求解的模型。

4.3.2 决策树

决策树是一种有监督的机器学习算法, 且对于分类和回归问题我们有 Decision Tree Classifier 和 Decision Tree Regressor 来解决对应的问题。决策树是一个基于树结构进行决策的方法，在每一个非叶子节点上我们都会对某一个特征进行属性测试，而样本根据结果进入到不同的子树当中，层层递进，直到最后进入叶子节点完成分类。

决定分类器学习性能的关键是我们该如何去制定每个非叶子节点上的特征划分，包括什么样的特征是应该用来选择的，这个特征又该何时放入决策树中呢 (即这个特征划分位于书中的哪一个非叶子节点) 等等问题。目前常用的算法有：ID3,C4.5 和 CART。

(1)ID3: ID3 算法是通过信息增益来选择特征递归地生成决策树。我们会对每一种特征都求出它对应的信息增益，然后选取信息增益最大对应的特征作为本次的分类依据。但是 ID3 算法也是有缺陷的。信息增益偏向取值较多的特征，这是由于当一种特征的取值过多时，更加容易取得纯度高的子集。

(2)C4.5: C4.5 算法可以算作 ID3 的升级，它的特征选择标准是信息增益比，即在原本信息增益的基础上增加了一个与特征取值正相关的分母，也因此在 ID3 中对取值较多特征的偏爱这一问题在 C4.5 中会得到较好的修正。

(3)CART: CART（Classification and Regression Tree）算法则把决策树变成了一个可以做回归任务的分类器。它的特征选择标准基尼指数。基尼指数表示在一个样本集合中随机选取的一个样本被分错的概率，因此，同信息熵一样，其值也是越小，样本集合越纯。通常我们让节点的取值为“是”和“否”，因此 CART 分类树是一棵递归的二叉树。

在决策树已经生成后，我们通常会采用剪枝的方式来减少其过拟合的程度。

决策树算法如下:

Algorithm 3 Decision Tree Algorithm [3]

Input: train set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
feature set $A = \{a_1, a_2, \ldots, a_d\}$.
Processing: Function :TreeGenerate(D, A)
Generate the Node;
if For all samples in D belongs to the category C **then**
Mark Node as a leaf node of category C; **return**
end if
if $A = \varnothing$ OR The samples in D take the same value on A **then**
Mark the Node as a leaf node and its category as the class with the largest number of samples in D; **return**
end if
Select the optimal division attribute a_* from A;
for Each of value a_*^v in a_* **do**
Generate a branch for Node;Let D_v denote the subset of samples in D that take the value a_*^v on a_*
if D_v is null **then**
Mark the branch node as a leaf node and its class as the class with the most samples in D; **return**
else
Take TreeGenerate($D_v, A \setminus \{a_*\}$) as a branch node
end if
end for
Output: A Decision Tree with Node as the root node

4.3.3 评价指标

与第二问类似，我们统计各个类别的样本数量，发现有一个类别样本数量很少:

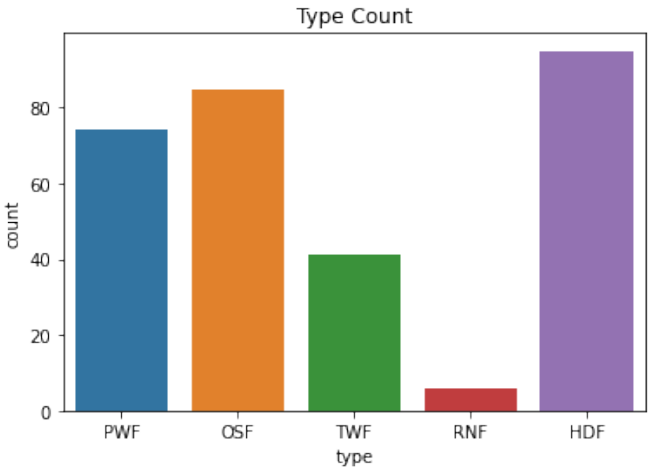


图 11: 各类别数量统计

因此，样本同样存在不平衡的问题，我们希望在预测时不漏某个类别，因此仍然采用 $F1$ 指标，不过我们采用另一个多分类指标：“$macro-F1$”，表达式如下：

$$
\begin{aligned}
macro-P &= \frac{1}{n}\sum_{i=1}^{n} P_i \\
macro-R &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
macro-F1 &= \frac{2}{\frac{1}{macro-P}+\frac{1}{macro-R}}
\end{aligned}
\tag{5}
$$

其中 n 为类别数。

4.3.4 问题三求解

首先我们对类别赋权重以平衡样本，然后我们选用决策树模型进行训练，使用网格搜索法进行自动调参，得到在最优参数情况下，5 折交叉验证时的 $macro-F1$ 为 0.8337，决策树参数如下：

属性	取值
class_weight	{3: 10}
criterion	entropy
max_depth	4
min_impurity_decrease	0.0
min_samples_leaf	1
splitter	best

表 3: 决策树参数

树模型如下所示：

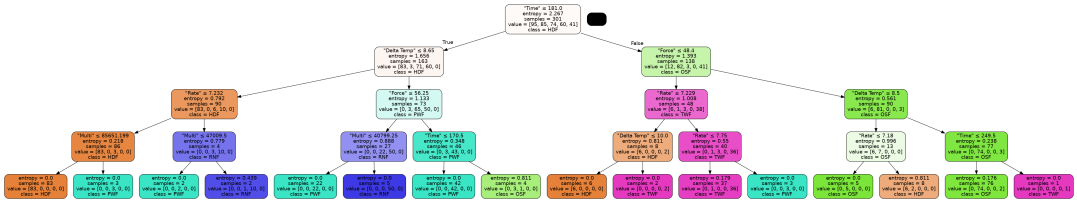


图 12: 决策树结果

详情请看附件。从决策树分类的结果我们可以看到“使用时长”、“转速”、“扭矩”、“温差”、“乘积”是最主要的几个因素，其余因素都被剪枝操作去除了。

4.4 问题四求解

在问题二和问题三中我们分别建立起了判别机械设备是否出问题的模型和在故障机械设备中进一步判别具体故障类型的模型。因此在对 1000 个数据进行预测的时候，我们会首先使用问题二中建立的模型对这些数据进行初步预测，将它们划分为正常和故障的两类，然后再使用在问题三中建立的模型进一步预测故障类的具体故障类别。

读取“forecast.xlsx”文件，将数据输入模型进行预测，一共找到 30 个异常样本，选取 10 个展示如下：

	id	code	level	in_temp	mac_temp	rate	force	time	label	type
746	9970	L47926	L	296.8	308.1	1289	62.0	199	1	OSF
442	9631	L47622	L	297.4	308.5	1399	61.5	61	1	PWF
993	9443	M18666	M	302.3	310.9	1360	44.0	67	1	HDF
995	9244	H33243	H	302.3	310.9	1366	48.4	130	1	HDF
405	9929	L47585	L	297.3	308.5	1350	57.6	186	1	OSF
50	9737	L47230	L	298.9	309.1	2861	4.6	143	1	PWF
998	9552	L51261	L	302.0	310.4	1336	58.2	110	1	HDF
194	9968	M15054	M	298.2	308.5	2678	10.7	86	1	PWF
463	9922	L47643	L	297.4	308.7	2874	4.2	118	1	PWF
603	9533	L47783	L	297.9	309.8	1336	71.6	31	1	PWF

图 13: 预测结果

4.5 问题五求解

在问题三中我们采用决策树来解决多分类问题，在决策树生成的过程中也是排列特征重要性的过程，我们将用在问题三中决策树模型的结果对各类故障选出明显相关的属性。

由于决策树的节点从上到下信息增益递减，因此我们选择能够使某一类成为主要类的第一个分叉节点作为主要影响节点，从中找到故障主要原因。得到 HDF、OSF、PWF、RNF、TWF 的主要特征属性为：“温差”、“扭矩”、“温差”、“转速扭矩乘积”、“扭矩”。我们对成因做如下解释：

- HDF 主要特征属性为温差小，查阅资料发现 [5]，发动机散热器故障时会出现漏水现象，水蒸发吸热导致设备温度降低，与室内温度的温差降低。
- OSF 主要特征属性为扭矩大，可解释为发动机过载故障时提供的动力更大，所以扭矩增大。
- PWF 主要特征属性为温差大，可解释为一些常见的电气事故，如短路等，由 $Q = I^2R$ 知发热量增加，导致温差增大。
- RNF 主要特征属性为转速扭矩乘积大，查阅资料发现 [6]，转速扭矩的乘积与功率近似成比例，这类故障可能是使得功率突然增大产生的故障，如系统误差发散等。
- TWF 主要特征属性为扭矩小，可解释为发动机磨损后转动更加困难，做功更多用于摩擦生热。

对于每一个类别，我们从根节点出发到达该类别对应的节点，形成一条轨迹，并从中选取绝大多数样本从属的轨迹作为规则，各类规则总结如下：

类别	规则
HDF	“时长”⩽181 and “温差”⩽8.65 and “转速”⩽7.23 and “转速扭矩乘积”⩽85651.2
OSF	“扭矩”⩽48.4 and “温差”⩽8.5 and 181<“时长”⩽249.5
PWF	(1)“时长”⩽181 and “温差”⩽8.65 and “扭矩”⩽56.25 and “转速扭矩乘积”⩽40799.25 (2)“时长”⩽170.5 and “温差”⩽8.65 and “扭矩”⩽56.25
RNF	“时长”⩽181 and “温差”⩽8.65 and “扭矩”⩽56.25
TWF	“时长”>181 and “扭矩”⩽48.4 and 7.23<“转速”⩽7.75

表 4: 各类故障潜在规则

其中转速是做了对数变换的结果。至此我们对潜在的规则进行了量化的分析。

5 模型评价

5.1 模型的优点

- 采用集成算法，将最优学习器进一步合并以追求更高的性能。
- 将特征选择和规则提取融入到模型训练中，一举多得。
- 模型能够处理变量之间的非线性关系，提高预测的准确率。

5.2 模型的缺点

- 模型数量多，调参难度高且更加耗时。
- 模型并不专门针对小样本检测任务，只能通过改变样本权重来平衡。

参考文献

[1] 周志华. 机器学习 [M]. 北京：清华大学出版社, 2016.1.184.

[2] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[C]// Knowledge Discovery and Data Mining. ACM, 2016.

[3] 李航. 统计学习方法 [M]. 清华大学出版社, 2012.235.

[4] 黄伟力, 黄伟建, 王飞, 等. 机械设备故障诊断技术及其发展趋势 [J]. 矿山机械, 2005, 33(1):3.

[5] 梁仕东. 发动机散热器溢水故障分析与排除 [J]. 建设机械技术与管理, 2013.

[6] 徐宝云, 陈民鉴. 发动机扭矩与转速关系经验公式试验研究 [J]. 车辆与动力技术, 1996.

附录

代码名称	代码解决的问题	代码电子版所在的位置
A1.ipynb	问题一的数据预处理、特征工程及特征选择	支撑材料中代码文件夹
A2.ipynb	问题二的模型训练与调参	支撑材料中代码文件夹
A3.ipynb	问题三的模型训练与调参	支撑材料中代码文件夹
A4.ipynb	问题四对附件数据进行预测	支撑材料中代码文件夹