离散制造过程中典型工件的质量符合率预测

突然乒通

团队简介

黄超，队长，华东师范大学软件工程2018级研究生。

王晨皓，队员，同济大学交通工程2010级学士，University of Alberta 交通工程2014级硕士，现任上海云砥信息科技有限公司部门经理。

邵柏潭，队员，江南大学物联网工程2016级本科生。

罗时超，队员，毕业于四川大学数据学院，现任新网银行风险建模工程师。

林宏涛，队员，华南理工大学自动化2017级研究生。

摘要

为了解决离散工件质量检测问题，我们首先对赛题数据进行了深入探索分析，得到5点具有重要指导意义的结论。继续深入研究P类特征与A类特征对结果的影响，确定了特征工程的方向，采用XGBoost Regressor预测A类特征，测试集中P9空值采用LightGBM Regressor预测填充。在模型构建阶段为提升方案的稳定性，最终模型采用5折交叉验证，并配合3个不同的随机数设置。最后对结果做后处理，生成提交结果。

关键词

特征工程，模型构建，交叉验证

1 赛题简介

1.1 赛题背景

在高端制造领域，随着数字化转型的深入推进，越来越多的数据可以被用来分析和学习，进而实现制造过程中重要决策和控制环节的智能化，例如生产质量管理。从数据驱动的方法来看，生产质量管理通常需要完成质量影响因素挖掘及质量预测、质量控制优化等环节，本赛题将关注于第一个环节，基于对潜在的相关参数及历史生产数据的分析，完成质量相关因素的确认和最终质量符合率的预测。在实际生产中，该环节的结果将是后续控制优化的重要依据。

1.2赛题任务

本赛题要求参赛者对给定的工艺参数组合所生产工件的质检标准符合率进行预测。

1.3赛题数据

本赛题提供数据中包含2类特征：(1) 工艺参数（Parameter）10项，表示生产工件的设备加工参数，以下称为P类特征，分别为P1, P2, …, P10；(2) 质量指标（Attribute）10项，表示产出工件的质量，以下称为A类特征,分别为A1, A2, …, A10。所有特征均经过脱敏处理。

复赛训练集包含12934个样本，测试集包含6000个样本，其中3000个样本的P类特征第9项（Parameter9）缺失，每50个样本分为一组，共120组。

测试集不包含A类特征。

本文采用复赛训练集与测试集进行说明与图表绘制。

1.4 评价指标

本赛题的预测目标为质检指标（不合格、合格、良、优），评价指标采用平均绝对误差（MAE）系数，计算方法如下：

其中 为预测样本， 为真实样本， 越接近1分数越高。

2 数据分析过程

本文提出的参赛建模策略基于以下数据分析中得出的结论，可总结为5点：

1. 特征中，P1~P4和A1~A3可视为连续型特征，P5~P10和A4~A10共13可视为离散型特征；
2. 训练集和测试集中部分特征部分取值的差异是由于保留位数不同造成；
3. 大部分特征经对数变换后呈典型正态分布形态，便于展开后续处理；
4. P类特征中，P1和P4在训练集和测试集上的分布存在差异，且在后续特征选择中发现P1~P4对模型的干扰较大；
5. A类特征中，A4~A10与预测目标均有显著关系，其中A4~A6特征对预测目标有决定性的影响。

以上结论对本文提出的预测方案均有重要指导意义。

2.1 特征的连续性与离散性

尽管P类特征与A类特征均以数值形式提供，在分析中发现，P类特征中P1~P4和A类特征中A1~A3的唯一值数量与样本数相同，可视为连续型特征；其他特征的唯一值数量均未超过样本数的10%，可视为离散型特征。

2.2 保留位数对P类特征的影响

基于章节2.1的结论，在比较特征在训练集和测试集上的分布时发现，离散型特征（P5~P10）中，部分特征的部分取值仅在训练集或验证集中出现（表1）。进一步探索后发现，这种情况大部分是由于某些取值的保留位数不同造成的。例如，对于P9特征，在训练集中存在取值0.004532889112547222，在测试集中存在取值0.0045328891125472216，但可以假设二者是相同的，其差异仅由导出时保留的位数不同造成。

表1：部分P类特征在训练集与测试集上取值的比较



2.3 对数变换对特征的影响

原始数据集中各项特征的取值范围较大，最小值一般接近于0而最大值超过10^5，且呈典型重尾分布形态。为了更好地分析特征，本队对所有特征均进行底数为2的对数变换，变换后的连续型特征和部分离散型特征呈典型正态分布。图1展示了一个典型特征（P1）在对数变换前后的分布直方图。

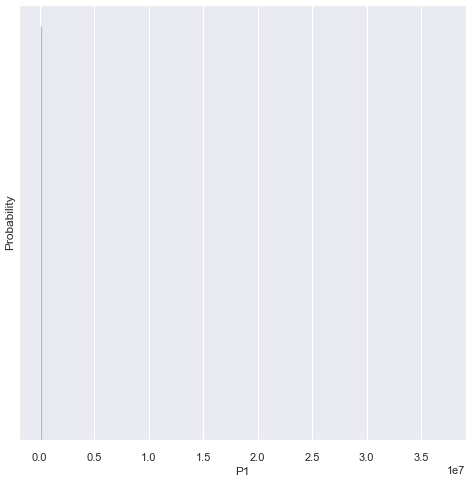
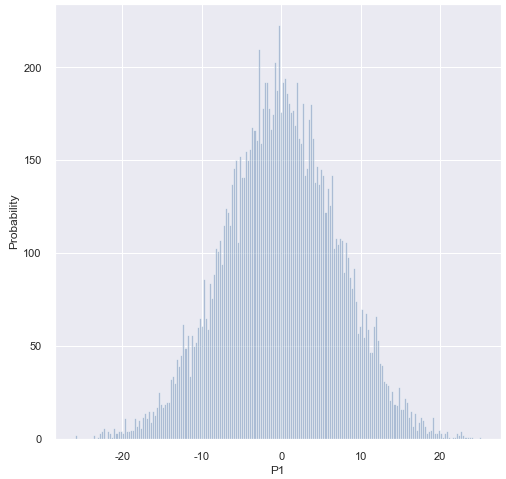
 

图1：对数变换前后P1特征的分布直方图

2.4 P类特征对结果的影响

分析中发现，P类特征中，经对数变换后的P1与P4在训练集和测试集上的分布存在差异，在训练集上的分布的方差较大。由于数据集规模较小，这种差异可能对结果造成影响。在实际线下验证中同时发现，当模型加入P1~P4特征时在验证集上的评分会下降，进一步利用 permutation feature importance 分析后发现，P1~P4由于取值数较多，对于基于树的机器学习模型而言更容易被选择作为分枝条件，在模型中的特征重要度较高，但在线下验证集中的重要度显著下降，即这4项特征对模型有一定的误导性。

2.5 A类特征对结果的影响

首先，除A1~A3外，其他A类特征对预测目标均有良好的区分度，显然是由于预测目标（工件质量评级）由工件质量指标综合得到的原因，而A1~A3特征是否加入对预测结果的影响较小。

其次，A4~A6特征对预测目标有非常显著的影响。图2展示了经对数变换后的A4, A5, A6特征构成的三维空间中，训练集样本的散点图，颜色代表样本质检指标。从图中可以发现：(1) 所有样本分布在两个相交的空间平面上；(2) 绝大部分样本在同一个平面，且样本的质检指标分布存在极为明显的规律。

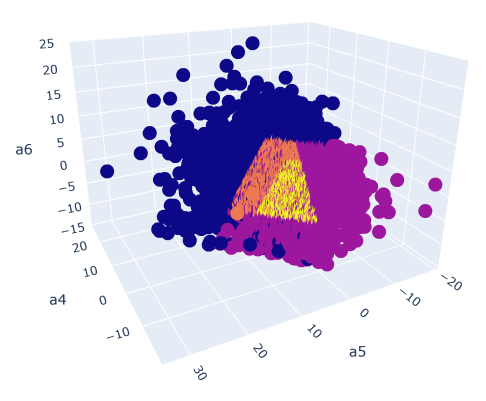
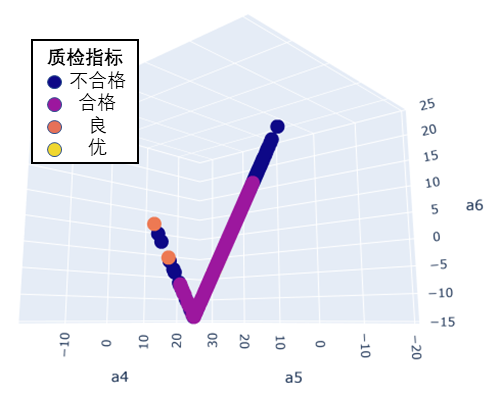


图2：训练集样本在A4~A6 特征的散点图

基于上述发现，对A4~A6特征做降维处理，得到降维后的特征K1, K2，得到如图4所示的二维平面散点图。图3结果充分说明，A4~A6特征对预测结果具有决定性的影响。在后续分析中，本队尝试将K1, K2与其他P类特征和A类特征叠加，但未再出现能进一步区分质检指标的组合。

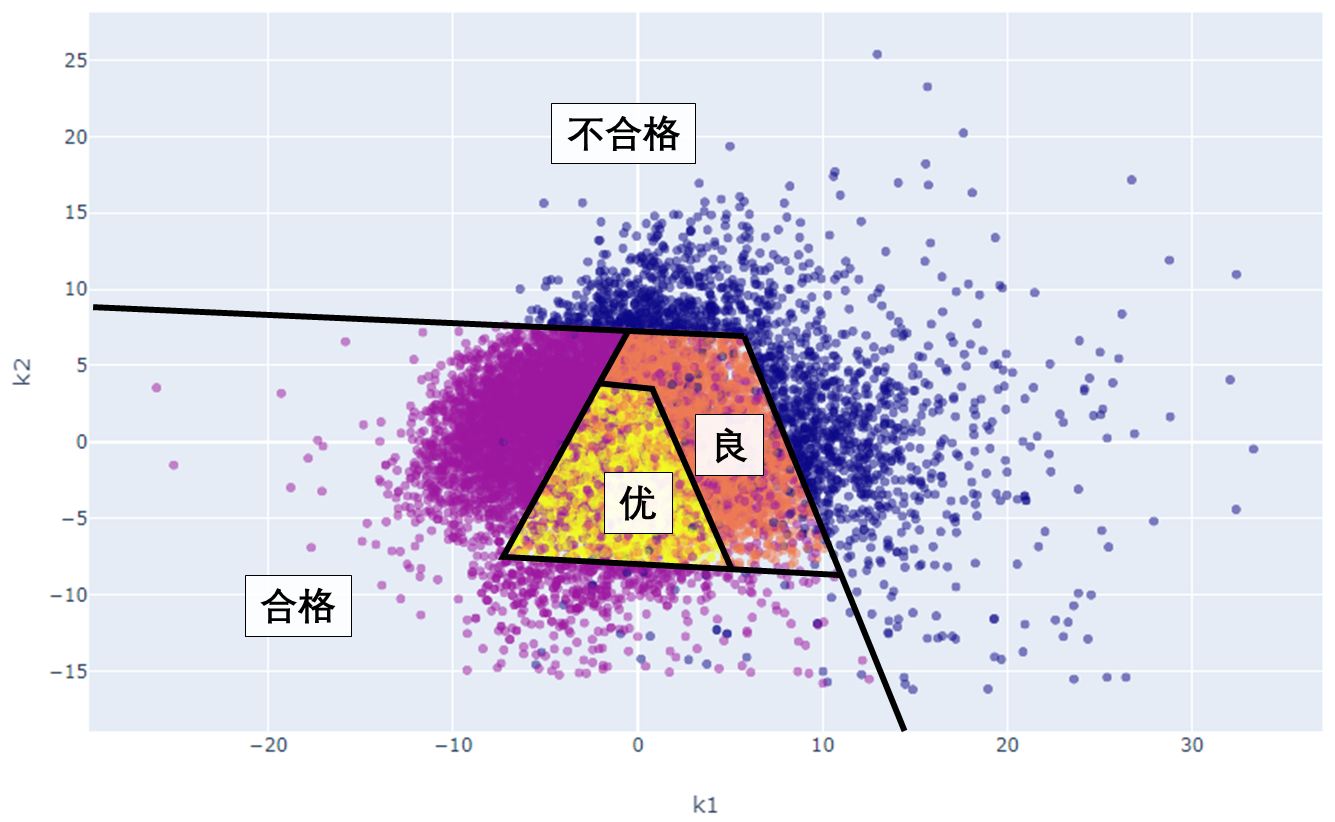


图3：训练集样本的 K1, K2 特征散点图

3 预处理，特征工程与建模

基于章节2中分析得到的结果，构建本队的建模方案，包含4项内容：数据预处理、特征工程、机器学习模型构建与结果后处理。

3.1 数据预处理

数据预处理环节包含以下2项操作：

(1) 对P类特征进行约简，即数值保留一定位数。若保留位数过小，可能导致部分接近的取值被合并，若过大则无法起到消除训练集和测试集差异的作用。经尝试，最终本队选择所有P类特征保留11位小数，处理后结果见表2，可与表1对比。

(2) 对P类特征和A类特征进行底数为2的对数变换，即 ，其中 为特征 的值。

表2：约简后P类特征在训练集与测试集上取值的比较



3.2 特征工程：P9 预测值

由于测试集中P9特征存在缺失值，且该特征在本队构建的模型中重要度较高，因此本队选择采用P5~P8和P10特征预测P9特征，并使用预测结果填充测试集中的缺失值。

根据章节2得到的结论，可将P9特征视为离散型特征，因此在预测前先对P9按数值升序排序进行重编码，编码后的P9特征的值为0, 1, …, 16，共17种取值。

经过对多种模型的尝试与比较，最终采用 LightGBM Regressor [1] 作为预测模型。

3.2 特征工程：A4, A5, A6 预测值

基于章节2中分析得到的结论，可确认A4~A6特征与预测目标有显著相关性。由于测试集不包含A类特征，因此考虑首先采用训练集P类特征分别预测A4~A6特征，再将预测得到的A4~A6特征作为训练集和测试集输入。具体操作如下：

1. 将训练集样本平均分为6份；
2. 每次使用5份训练集数据，以P5~P10特征为输入，建立模型预测A4~A6特征，再用训练得到的模型预测剩余1份训练集数据和测试集数据的A4~A6特征；
3. 重复第 (2) 步6次，得到所有训练集数据的A4~A6 特征预测值，以及测试集数据的A4~A6特征预测值的均值。

经过对多种模型的尝试与比较，最终采用 XGBoost Regressor [2] 作为预测模型。

3.4 特征工程：其他统计特征

本队在比赛期间共尝试构建106个统计特征，经特征选择后，最终方案中保留以下2类共18个统计特征：

(1) 以一项特征作为分组条件，对另一项特征计算分组后的平均值、标准差、总和、唯一值计数，最终保留15项该类特征；

(2) 频率编码，最终保留3项该类特征。

3.5 特征工程：特征选择

基于章节2得出的结论，最终作为模型输入的特征为：P类特征中的P5~P10，A类特征中的A4~A6（预测值），以及章节3.4中提到的统计特征。

3.6 模型构建

经多次线下与线上验证后发现，在采用不同随机数时结果评分存在较明显的差异。为提升最终方案的稳定性，最终模型采用5折交叉验证，并配合3个不同的随机数设置，即3次5折交叉验证，最终得到15次预测值，将其平均值作为最终预测结果，能取得相对稳定的线上评分。

经多次尝试后，最终采用 XGBoost Classifier 作为最终模型。

3.7 结果后处理

根据赛题中的描述，在实际生产中，同一组工艺参数设定下生产的工件会出现多种质检结果，因此最终分类器模型的输出结果采用各分类的预测概率，则测试集中单组的预测结果即为组内所有样本的分类预测概率之和。

4 方案效果

本文中提出的预测方案在比赛的复核赛B榜线上评分为0.7079336，位列第一名。

致谢

感谢CCF大数据与计算智能大赛和DataFountain平台给我们提供了一个近距离接触人工智能应用的宝贵的机会。感谢活跃在DataFountain这个平台的数据科学家以及工作人员，你们的无私奉献促进发展了国内整个数据科学竞赛的水平，也让更多年轻人能够有机会感受人工智能的魅力。我们队在这次比赛中学习到了很多，我们能有这样的成绩离不开你们的帮助，在此致以我们最诚挚的谢意和最衷心的感谢。

参考

[1] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]//Advances in Neural Information Processing Systems. 2017: 3146-3154.

[2] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016: 785-794.