# 离散制造过程中典型工件的质量符合率预测

孙浩然[1] 17182627 刘超 17182611[1] 石宇 17182610[1]

[1]（北京航空航天大学北京学院 北京 100191）

# Prediction of Quality Consistency of Typical Workpieces in Discrete Manufacturing Process

Sun Haoran[1] , Liu Chao[1] , and Shi Yu[1]

[1]（*College of Beijing, Beihang University, Beijing* 100191）

**Abstract** In the actual situation, it is more practical to predict the compliance rate of quality inspection standards. It is more practical to predict the quality inspection of each workpiece. Through the analysis of the data in the training set, we find that the P and A special features are discrete features. We Finally, the features of the input model are P5 ~ P10 with P features, and CatBoost Classifier is selected as the final model. A 5-fold cross-validation was performed five times to obtain the training and test set prediction results. Because the label of our test is contained in the train, we merge the test into the train to facilitate the subsequent value retrieval and generate the data set. We use 50% cross-validation and set 5 random numbers as random number parameters. In order to reduce the impact of random numbers, take the average. These random number parameters are used to randomly select 4/5 of the original training set as the training set and 1/5 of the data as the test set. Finally, the results of different random numbers are averaged 5 times to obtain the final prediction result, and the prediction success rate of the training set is output.

**Key words** quality inspection compliance rate；CatBoost Classifier model; 5 fold cross validation

**摘要** 在实际情况中，预测质检合格标准符合率相比较预测各个工件的质检结果会更加具有实际意义，通过对训练集中数据的画图分析，我们发现 P 类和 A 类特征的一部分是离散的特征，我们最终选择输入模型的特征为 P 类特征的 P5~P10，选择 CatBoost Classifier 作为最终模型。采用 5 折交叉验证，进行 5 次，得到训练集和测试集预测结果。因为我们的 test 的标签是包含于 train 内的，所以我们将 test 合并到 train 中便于后边取值，生成了 data 集。我们采用 5 折交叉验证并设 5 个随机数为随机数参数，为了减少随机数的影响取平均值。这些随机数参数用于随机选取原本训练集中 4/5 的数据作为训练集，1/5 的数据作为测试集。最后，将 5 次采用不同随机数的结果取平均值得到最终预测结果，输出训练集的预测成功率。

**关键词** 质检合格标准符合率；CatBoost Classifier 模型；5 折交叉验证

**中图法分类号** TP391

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，本文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. 赛题简介

### 1.1 赛题任务

由于在实际生产中，同一组工艺参数设定下生产的工件会出现多种质检结果，所以我们针对各组工艺参数定义其质检标准符合率，即为该组工艺参数生产的工件的质检结果分别符合优、良、合格与不合格四类指标的比率。相比预测各个工件的质检结果，预测

该质检标准符合率会更具有实际意义。

本赛题要求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。

### 1.2 赛题数据

在此任务中，以某典型工件生产过程为例，提供一系列数据。该数据来源于某工厂采集的真实数据，已做脱敏处理。

（1）训练数据将提供：

A：工艺参数（如设备加工参数）

B：工件的质量数据

C：工件所符合的质检指标

（2）测试数据将提供：

A：工艺参数（如设备加工参数）

这些数据中包含两类特征：工艺参数（parameter）共 10 项，表示工件的加工参数，以下简称为 P 类特征；质量数据（attribute）共 10 项，表示产出工件的质量，以下简称为 A 类特征。

### 1.3 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 MAE 系数，计算方法如下：

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|pred_i - y_i|$$

$$\mathrm{Score} = \frac{1}{1 + 10 * MAE}$$

其中 $\mathrm{pred}_i$ 为预测样本，$y_i$ 为真实样本。最终结果越接近 1 分数越高。

## 2 模型构建

### 2.1 数据分析

通过对 first_round_training_data.csv 中数据的画图分析（图 1），可以发现 P 类 5-10 特征和 A 类 4-10 特征是离散的特征[1]。

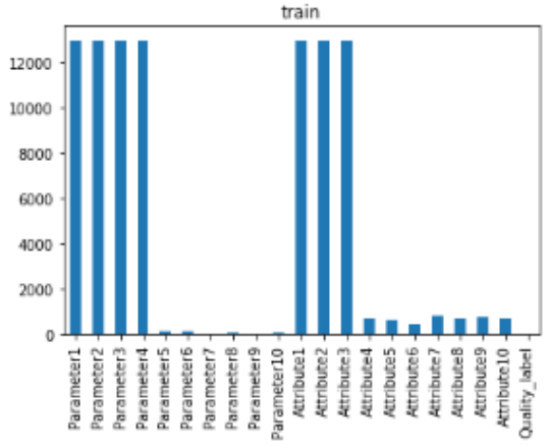



图 1

P 类特征中，P1 和 P4 在训练集和测试集上的分布存在差异，而且在后续选择中发现 P1~P4 对模型影响较大。

### 2.2 P 类特征对结果的影响

分析中发现，当模型加入 P1~P4 特征时，在验证集上的分数会下降。由于这四个特征取值数较多，对于基于树的机器学习模型而言更容易被选择为分枝条件，这四个特征对模型有一定的误导性。

### 2.3 特征选择

基于上述分析，最终选择输入模型的特征为 P 类特征的 P5~P10。

### 2.4 模型构建

最终选择 CatBoost Classifier 作为最终模型。采用 5 折交叉验证，进行 5 次，得到训练集和测试集预测结果。由于使用 5 个随机数作为划分参数，故每次预测值占比为 1/5，最终输出预测结果。

### 2.5 CatBoost 简介

我们可以使用 CatBoost，而不需要任何显式的预处理来将类别转换为数字。CatBoost 使用在各种统计上的分类特征和数值特征的组合将分类值转换成数字。

它减少了对广泛的超参数调优的需求，并降低了过度拟合的几率，这也导致了模型变得更加具有通用性。但它还包含一些参数，比如树的数量、学习速率、正则化、树的深度等等[2]。

将 XGBoost、LightGBM 和 CatBoost 相比较，会发现 CatBoost 在大多数情况下的 log-loss 是最低的。说明了 CatBoost 对调优和默认模型的性能都更好。而且，CatBoost 不需要像 XGBoost 和 LightGBM 那样将数据集转换为任何特定格式。

### 2.6 结果输出

| Group | Excellent r | Good ratio | Pass ratio | Fail ratio |
|---|---|---|---|---|
| 0 | 0.221472 | 0.33038 | 0.256862 | 0.191285 |
| 1 | 0.189062 | 0.269364 | 0.369364 | 0.172209 |
| 2 | 0.220332 | 0.308625 | 0.299373 | 0.171671 |
| 3 | 0.241119 | 0.295667 | 0.349292 | 0.113922 |
| 4 | 0.179301 | 0.367642 | 0.25648 | 0.196578 |
| 5 | 0.189673 | 0.263501 | 0.386863 | 0.159963 |

图 2

根据题意，最终分类器模型输出的是采用各分类的预测概率（图 2）。

### 2.7 反思

最终网站上给出了排名前五的大佬的思路，经过仔细阅读学习后发现了我们的很多不足。

（1）A 类特征中的 A4~A6 特征对预测目标有非常显著的影响，因为测试集中不包含 A 类特征，所以可以通过特征 P5~P10 来预测这三个特征。测试集中不包含 A 类特征，也是我们没有认真挖掘 A 类特征信息的原因，导致我们最终没有用到 A 类特征。下次若再有机会参赛，相信我们不会再出这种问题。[3]

（2）我们没有对 P 类特征进行数据预处理，而 P 类数据的数值位数是不一致的，如果能保留一个合适的位数，可以消除一些训练集和测试集的差异。

## 3 数据的处理与代码的实现

在本节中，我们将具体介绍我们是如何基于 python 实现上述模型的。

### 3.1 预处理

首先，我们用 pandas 包中的 csv 文件处理函数，读取了训练集和测试集的内容（图 3）。

```
train = pd.read_csv('first_round_training_data.csv')
test = pd.read_csv('first_round_testing_data.csv')
```

图 3

因为我们的 test 的标签是包含于 train 内的，所以我们将 test 合并到 train 中便于后边取值，生成了 data 集（图 4）。

```
data = train.append(test).reset_index(drop=True)
```

图 4

由于我们只需要特征 Parameter5- Parameter10，下面我们只取这些特征的值和已有的结果作为新的训练集和测试集（图 5），其中 feature_name 即为标签 Parameter5- Parameter10。

```
X_train = data[tr_index][feature_name].reset_index(drop=True)
y = data[tr_index]['label'].reset_index(drop=True).astype(int)
X_test = data[~tr_index][feature_name].reset_index(drop=True)
```

图 5

至此，我们的预处理部分结束。

### 3.2 模型构建

根据第二节的思路，我们现在需要采用 5 折交叉验证，并设 5 个随机数为随机数参数，最外层循环次数设为 5（图 6），为了减少随机数的影响取平均值。

```
seeds = [19970412, 2019 * 2 + 1024, 4096, 2048, 1024]
num_model_seed = 5
```

图 6

这些随机数参数用于随机选取原本训练集中 4/5 的数据作为训练集，1/5 的数据作为测试集（图 7）。

```
skf = StratifiedKFold(n_splits=5, random_state=seeds[model_seed], shuffle=True)
for index, (train_index, test_index) in enumerate(skf.split(X_train, y)):
```

图 7

用每次选取的训练集训练模型，再将模型用于预测测试集和上述每次分离的 1/5 训练集，由于测试集未被拆分，所以测试集每次预测结果权重为 1/5，而循环之后原来训练集则已经被全部预测，用于计算准确度，测试集预测结果为 5 次交叉验证结果的均值（图 8）。

```
cbt_model = cbt.CatBoostClassifier(iterations=800,learning_rate=0.01
cbt_model.fit(train_x, train_y ,eval_set=(train_x, train_y))
oof_cat[test_index] += cbt_model.predict_proba(test_x)
prediction_cat += cbt_model.predict_proba(X_test)/5
```

图 8

最后，将 5 次采用不同随机数的结果取平均值得到最终预测结果，输出训练集的预测成功率，并把测试集结果转成 csv 文件。

## 4 总结与致谢

此次竞赛，我们小队取得了预赛 A 榜 533 名，B 榜 414 名的成绩。这个成绩固然不是很行，但是对于初次接触数据挖掘的我们来说，却是实实在在的见识到了很多新东西，学到了很多新知识，也在参考大佬的思路中学到了一些套路。虽然成绩不理想，也没能进入复赛，但是我们收获良多。

在此，也感谢竞赛中各位大佬分享的思路，让我们学习到了很多。更感谢王静远老师的授课，教会了我们很多数据挖掘方面的知识，并让我们参与这次大赛，学会了很多东西。同时也感谢一直在为这门课操劳的助教们。非常感谢！

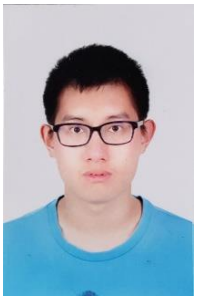
## 参 考 文 献

[1] https://discussion.datafountain.cn/questions/2234/answers/23334

[2] http://www.atyun.com/4650.html

[3] https://discussion.datafountain.cn/questions/2234/answers/23331

**Sun Haoran** born in 1999. Computer Science and Technology.



**Liu Chao,** born in 1999. Computer Science and Technology

**Shi Yu,** born in 1999. Computer Science and Technology.

.