**离散制造过程中典型工件的质量符合率预测方法**

尹子鑫1 付佳辉1 吴桐雨1

1（北京航空航天大学计算机学院 北京 100191）

1（yzx835@buaa.edu.cn）

#### Prediction method of quality conformity rate of typical workpiece in discrete manufacturing process

Yin Zixin1 , Fu Jiahui1, and Wu Tongyu1

1（*School of Computer Science and Engineering*, *Beihang University*, *Beijing* 100191）

**Abstract** In the field of high-end manufacturing, with the deepening of digital transformation, more and more data can be used for analysis and learning, so as to realize the intelligence of important decision-making and control links in the manufacturing process. For the manufacturing process of discrete workpieces, based on the analysis of potential relevant parameters and historical production data, this paper completes the confirmation of quality related factors and the final quality compliance rate Prediction. In the actual production, the result of this link will be an important basis for the follow-up control optimization. In this paper, based on the historical manufacturing parameter data of the workpiece, firstly, the data analysis is carried out, some useless features are deleted according to the data distribution features, secondly, the features with large correlation coefficient are fused based on the feature engineering, then the boosting model is selected for training and optimization, finally, the model fusion method is used to further improve the accuracy of prediction.

**Key words** discrete process, workpiece manufacturing, boosting, model fusion

摘要 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，本文针对离散工件的制造过程，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。本文基于工件的历史制造参数数据，首先进行数据分析工作，根据数据分布特征删除了部分无用特征，其次基于特征工程融合相关系数较大的特征，然后选用boosting模型进行训练调优，最后采用模型融合的方法进一步提高了预测的准确性。

关键词 离散过程，工件制造，boosting，模型融合

本文基于工件的历史制造参数数据，首先进行数据分析工作，根据数据分布特征删除了部分无用特征，其次基于特征工程融合相关系数较大的特征，然后选用boosting模型进行训练调优，最后采用模型融合的方法进一步提高了预测的准确性。

## 1 问题分析

训练数据的形式是(Parameter, Attribute, Label)，测试数据的形式是(Parameter)，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，将预测该质检标准符合率作为目标对生产的帮助会更具有实际意义。

需要注意的是，上述数据的产生并不是给定一组生产参数Parameter，生产出一个工件，测量其工件属性Attribute，然后根据工业界的标准确定该工件的指标Label。而是对每一组生产参数Parameter，均生产若干个工件，在理想情况下该批工件的属性Attribute应该服从某个分布，然后从该批工件中随机抽取一个样本，将该样本的Attribute和Label作为训练集中对应的数据。所以我们的任务并不是预测某个实际的工件的质检结果，而是根据生产参数Parameter来预测工件属性Attribute的一个分布特征，从而预测出质检标准符合率。

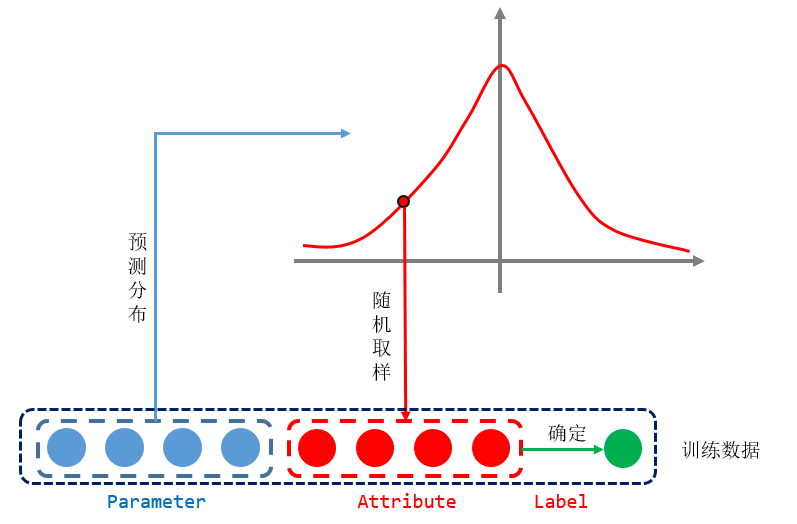


Fig.1 Generation of training data.

图1 训练数据的产生

所以对该问题，初始的想法有两条思路，一种是根据生产参数Parameter直接预测Label的分布，该方法简单直接，可以用作衡量模型好坏的baseline；另一种是根据生产参数Parameter预测工件质量Attribute的分布，再由Attribute得到Label，该方法更符合工业生产的实际情况，因为由一个工件属性得到工件的质检指标应该是一个多维线性可分的分类模型，可以很好的使用树模型刻画，只要对Attribute的回归预测足够准确，该方法应该有较高的准确率。



Fig. 2 Classification method.

图2 分类方法



Fig. 3 Regression and classification method.

图3 回归加分类方法

**2数据预处理**

**2. 1数据分析**

对训练数据在不同Label下的各个Parameter特征的分布情况进行初步的分析，例如下图：

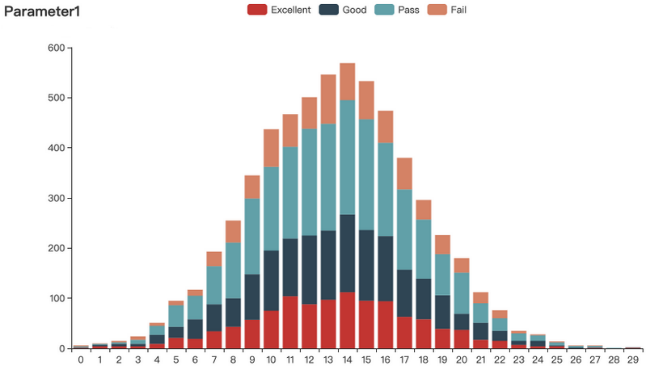


Fig. 4 Distribution of Parameter1.

图4 Parameter1的分布

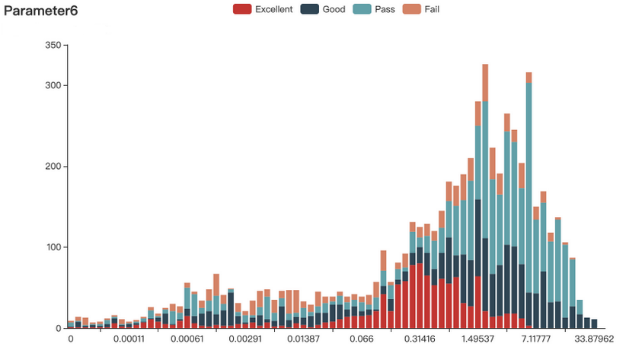


Fig. 5 Distribution of Parameter6.

图5 Parameter6的分布

在此分析的基础上，可以初步把Parameter特征分为两类：

1. Parameter1-4：该类别的特征是在每个Label中都近似于服从正太分布形式，数据的范围较小，分布比较均匀，在每个Label下的分布基本一致。对于这类特征，由于其分布一致从而不会给分类带来有用的信息，所以将其从模型的输入中删除。

2. Parameter5-10：该类别的特征是数据在每个Label上的分布差异较大，数据本身的数量级差异较大，分布极其不均匀，类似于长尾分布的特征。所以对其做对数处理，缩小其存在的范围，以便作为模型的输入。



Fig. 6 Make log changes to class 2 Parameters.

图6对2类Parameter做log变化

**2. 2特征工程**

尝试基于原始的Parameter特征，寻找其中相关性较强的特征组合，进行线性组合尝试构建新的特征作为输入。

使用皮尔逊相关系数衡量两个变量之间的相关程度：

，(1)

基于训练集数据我们得到如下结果：

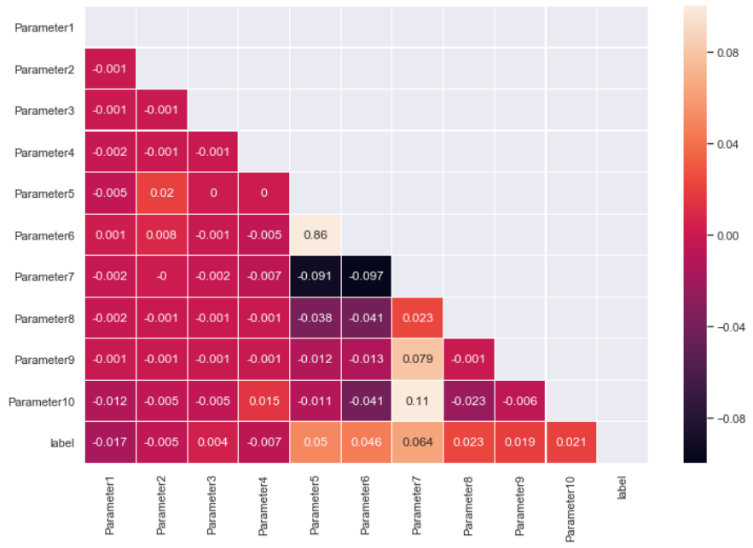


Fig. 7 Parameter correlation calculation.

图7 Parameter相关性计算

从分析结果我们可以知道除了Parameter5和Parameter6之间的相关性较高之外，其余Parameter之间的相关程度都比较低，所以将Parameter5+Parameter6作为一个新的特征添加到输入中去。

**2. 3数据可视化**

工件最后的Label是由Attribute属性简单的线性决定的，所以使用Attribute通过PCA降维之后在三维空间中展现Label的分布情况，具体结果如下：

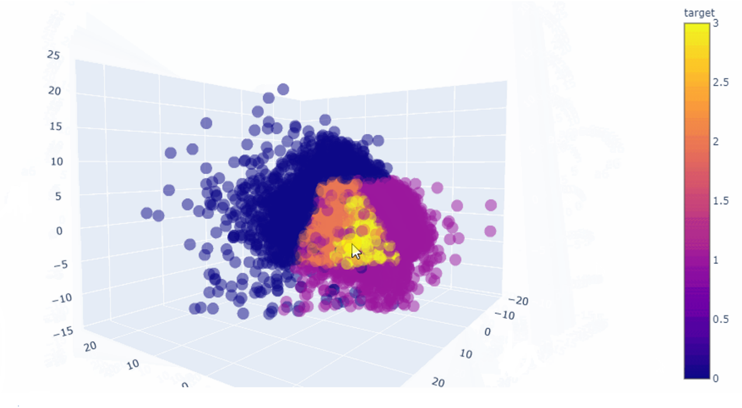


Fig. 8 Label distribution in 3D.

图8 三维中的Label分布

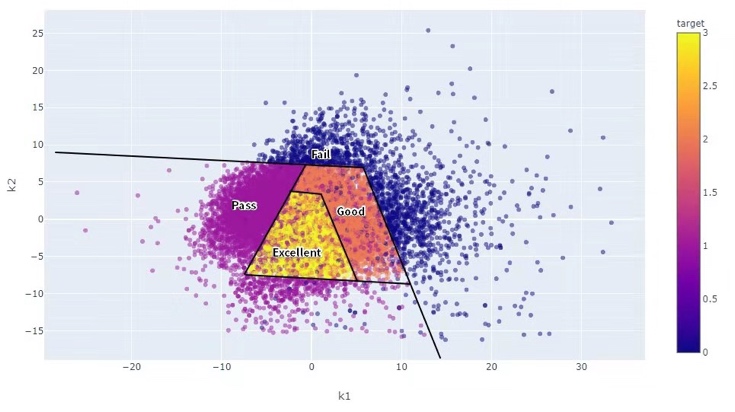


Fig. 9 Label distribution on section.

图9 截面上的Label分布

上述的在二维界面的可视化中不同Label之间的分割界限十分明显，并且界限都是线性函数可以描述的，说明了依靠Attribute可以很好的将Label线性分割，进一步证明了我们的第二种思路的正确性。

**3模型训练**

我们小组尝试了多种方法，我们最先使用了一个简单的神经网络，但是效果不尽人意；后来我们尝试了catboost、lightgbm等树模型，效果有了明显的提升；最后我们又尝试使用了多种模型融合的方法。

**3. 1神经网络**

我们最初尝试了神经网络的做法，模型是一个非常简单的5层线性神经网络。我们对Parameters数据取了log后并作了归一化处理，作为模型的输入。使用模型做分类任务，使用交叉熵作为损失函数，最后模型的softmax层的输出作为结果的概率分布。

我们在此基础上更改参数，做了多次尝试。我们尝试改变网络层数，改变模型学习率和训练轮数，尝试对输入数据做离群点数据去除。最终最好的结果在为0.4800（初赛）。

**3. 2 Catboost分类**

参考了网络上他人的baseline分享，才了解到catboost这个强大的模型。通过研究和学习catboost的官方文档，我们组又尝试使用catboost来直接使用Parameters做分类，来预测结果分布。由于catboost的强大功能，我们可以直接将选择好的特征输入模型，就可以学习出结果。

我们多次尝试更改训练轮数、学习率、树模型的其他参数，最终最好结果为0.6913（初赛）。

**3. 3 Catboost回归加分类**

我们又参考了网络上另一个baseline的分享，那个baseline使用的是先使用Parameters回归出Attributes，然后使用Parameters与预测出来的Attributes一起来做分类，预测结果分布。我们基于这个思想，使用catboost先用Parameter5~10预测出Attribute1到Attribute10，然后使用初始的Parameter5~10和预测到的Attribute1~10，预测最终结果。

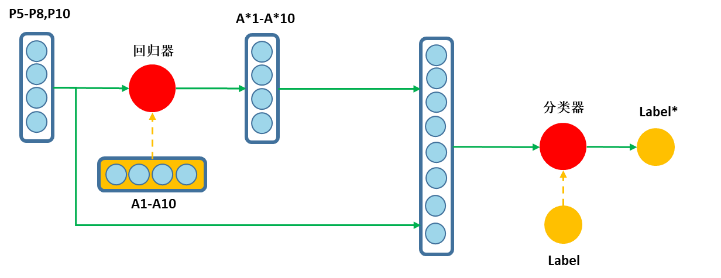


Fig. 10 Catboost regression and classification.

图10 catboost回归加分类模型

在尝试多种参数之后，最终最好结果为0.7005（初赛）。

**3. 4 模型融合**

在复赛时，由于之前使用的模型效果相对不是很好，所以我们尝试使用模型融合方法。回归任务使用catboost、lightgbm、xgboost同时做回归，然后将结果按照60%，20%，20%的权值加和。分类使用catboost、lightgbm，结果按照75%，25%的权重加和。权重的选取，我是考虑了单个模型的效果，如果某个模型的效果比较好，就给这个模型一个比较大的权重。最终在复赛的到了0.6987（b榜）的成绩。

**Table 1 The Results on Models**

**表1 各个模型的最终得分**

|  |  |
| --- | --- |
| *Model* | *Result* |
| 神经网络 | 0.4800（初赛a榜） |
| Catboost分类 | 0.6913（初赛a榜） |
| Catboost回归加分类 | 0.7005（初赛a榜）0.6279（复赛a榜） |
| 模型融合 | 0.6687（复赛a榜）0.6987（复赛b榜） |

**6** 总结

本文基于工件的历史制造参数数据，首先进行数据分析工作，根据数据分布特征删除了部分无用特征，其次基于特征工程融合相关系数较大的特征，然后选用boosting模型进行训练调优，最后采用模型融合的方法进一步提高了预测的准确性。

通过本次比赛，我们在学习了一些他人的解决方案之后，发现许多优秀队伍在模型特征选择时，花费的心思是最多的。有了更好，更能反应样本信息的特征要比训练一个模型来学习这些特征要高效的多，最终结果也会更好。