哈爾濱工業大學

HARBIN INSTITUTE OF TECHNOLOGY

2020 年 春 季学期本科生课程考核

（读书报告、研究报告）

考 核 科 目	人 工 智 能 与 机 器 学 习		
学生所在院（系）	计 算 机 科 学 与 技 术		
学 生 所 在 专 业	大 数 据 科 学 与 技 术		
学 生 姓 名	彭 钰 驯		
学 号	1 1 7 0 3 0 0 9 1 6		
考 核 结 果		阅 卷 人	李 永 立

摘　要

关键词：multi-class　随机森林　梯度提升树　神经网络

摘要：　本文来自大数据竞赛网站，离散制造过程中典型工件的质量符合率预测问题，属于 multi-class 问题。本文按照经典数据挖掘流程进行，依次进行了探索性数据分析（EDA），特征工程，模型训练与调优，模型融合。探索性数据分析阶段多角度进行了参数选取；模型训练与调优阶段采用了随机森林，梯度提升树和神经网络三种模型，提高了预测的准确度和可解释性。

目 录

第 1 章 绪　论

1.1 研究问题的背景

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，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。

该问题来自于大数据竞赛网站 DataFountain，旨在对离散制造过程中工件的质量进行符合率预测。在此任务中，以某典型工件生产过程为例，赛题将提供给参赛者一系列工艺参数，以及在相应工艺参数下所生产工件的质量数据。该数据来源于某工厂采集的真实数据，已做脱敏处理。数据分为 A、B、C 三类，要求参赛者基于 A 类数据计算出各组工艺参数的工件质检标准符合率。赛题评价指标使用 MAE 系数对网站的测试集（无分类结果）进行预测，但是由于网站在线评测已关闭，原本的测试集无法使用，而只用训练集无法用 MAE 作为指标（给出的训练集和测试集结构不一样），因此改用多分类准确率作为指标，只用赛题的训练集。

1.2 研究问题的挑战

Multi-class 问题是一个在工业生产中广泛存在的问题，它不同于图像的 multi-label 问题上近年来提出了很多高效的模型，如 VGG 等，也不像自然语言处理的 multi-label 中可以用 BERT 等提高准确率。在非图像处理的 multi-class 问题中分类准确率高的方法十分有限，特别是合适的神经网络模型。

其次，由于本题是一道大数据竞赛题，在参数选取和特征工程上也有一定的挑战，由于没有上过大数据挖掘的相关课程，一开始研究问题时对于数据挖掘的一般过程和相关方法还不太了解，也缺乏搭建神经网络的经验。

1.3 当前研究工作的不足之处

对于多分类问题的神经网络模型还有待优化，目前采取的 DNN 只取得了接近随机森林和梯度提升树的分类准确率，模型结构还有待进一步优化。还可以用 bagging 的方式将随机森林、梯度提升树、神经网络三个模型融合，三个模型得出的结果投票，得票多的作为预测的结果（由于神经网络模型和树模型在建立过程中

对类别变量转化为数值变量采用了不同的方式，本文未进行投票）。

1.4 本文的工作要解决的问题以及方法

本文要解决的问题是根据赛题所给的数据，训练多分类模型，对工件质量进行预测，即将工件根据参数进行多分类。本文的处理方法为先合并赛题数据，进行探索性数据分析，选取合适的属性，再进行特征工程，将类别变量（categorical function）转化为数值变量（神经网络中采用了 one-hot 编码，随机森林和梯度提升树模型中采用了 factorize）。分别用了随机森林、梯度提升树、神经网络训练多分类模型，经过模型调优后进行分类预测。

1.5 本文的贡献

对比了随机森林，梯度提升树和神经网络三种方法在多分类问题上的使用结果，探索了对于多分类问题各种神经网络的效果。采用数据挖掘的方法，对赛题数据进行了探索性数据分析以选取合适的属性，进行特征工程利用两种方法将类别变量（categorical function）转化为数值变量。

1.6 章节安排

本文依据大数据竞赛的流程依次介绍探索性数据分析，特征工程，模型训练与调优，模型融合四个部分。之后展示了所做的对比性实验，最后得出结论。

第 2 章 系统/方法框架

2.1 系统框架

2.1.1 梯度提升树：Catboost

CatBoost（categorical boosting）是一种能够很好地处理类别型特征的梯度提升算法库。它自动采用特殊的方式处理类别型特征（categorical features）。首先对 categorical features 做一些统计，计算某个类别特征（category）出现的频率，之后加上超参数，生成新的数值型特征（numerical features）。catboost 还使用了组合类别特征，可以利用到特征之间的联系，这极大的丰富了特征维度。catboost 的基模型采用的是对称树，同时计算 leaf-value 方式和传统的 boosting 算法也不一样，传统的 boosting 算法计算的是平均数，而 catboost 在这方面做了优化采用了其他的算法，这些改进都能防止模型过拟合。

2.1.2 神经网络：Keras

Keras 是一个用 Python 编写的高级神经网络 API，它能够以 TensorFlow, CNTK, 或者 Theano 作为后端运行（在本题中我采用的是 tensorflow 作为后端）。它能将 idea 迅速转换为结果，适合于小型环境(实验室、数据竞赛)。相比于传统的神经网络编写方式，它模块化程度更高，更易扩展，对用户也更友好，降低了神经网络的搭建难度，使得我们可以更集中于神经网络模型本身。

2.2 各部分简介

本文依据大数据竞赛的流程依次介绍探索性数据分析，特征工程，模型训练与调优，模型融合四个部分。之后展示了所做的对比性实验，最后得出结论。

第 3 章 探索性数据分析

3.1 赛题数据

合并第一轮和第二轮的数据集，根据题目要求对 parameter1-10 进行分析。赛题数据包括三种类型的字段 A，B，C。其中 A 类字段为十个工艺参数 parameter1-10，数据类型为 Float；B 类字段为 10 个工件属性，数据类型为 Float；C 类字段为工件所符合的质检指标 Quality_label，有四个类型：Excellent，Good，Pass，Fail。数据无缺失值，数据均经过了脱敏处理。赛题任务要求利用 A 类字段的工艺参数，预测出工件的质检指标。

数据：https://www.datafountain.cn/competitions/351/datasets

3.2 数据分布

工艺参数包括十个字段 Parameter1-10，且各个字段的数量级差异很大。通过 EDA，我们可以发现十种工艺参数按照数据的分布情况大体上可以分为两类。第一类数据为 Parameter1-4，每个值都不重复；第二类为 Parameter5-10。以 Parameter1 和 Parameter5 为例（其他类似），数据分布如下图所示：

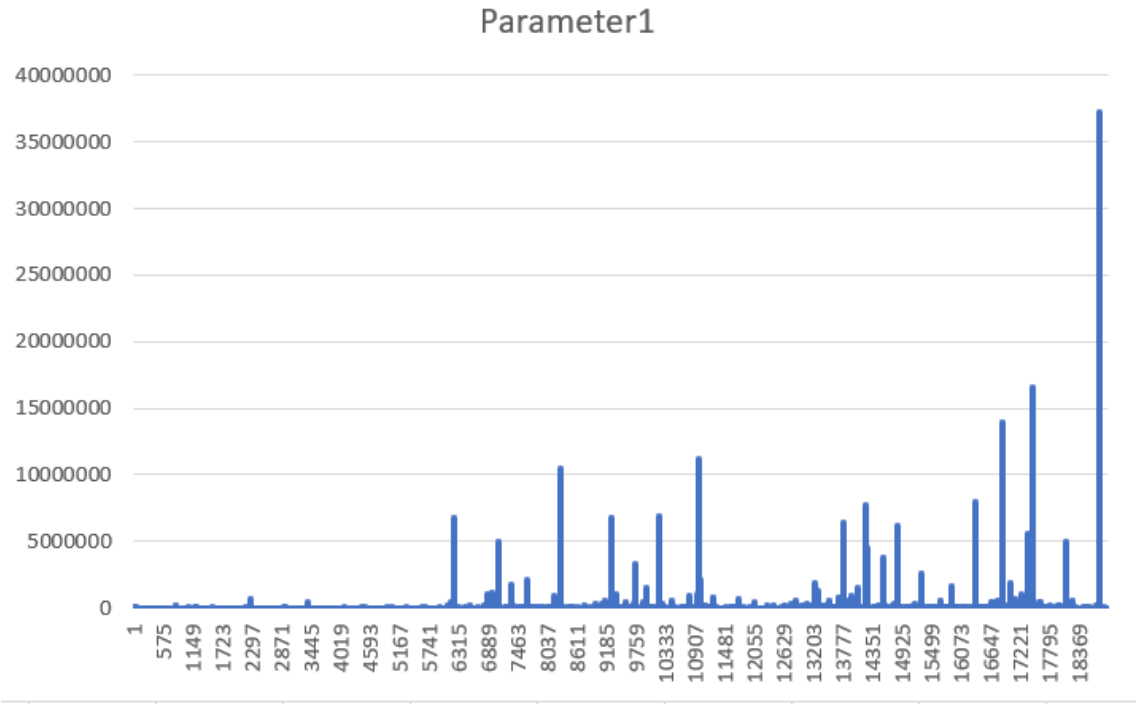

图 1

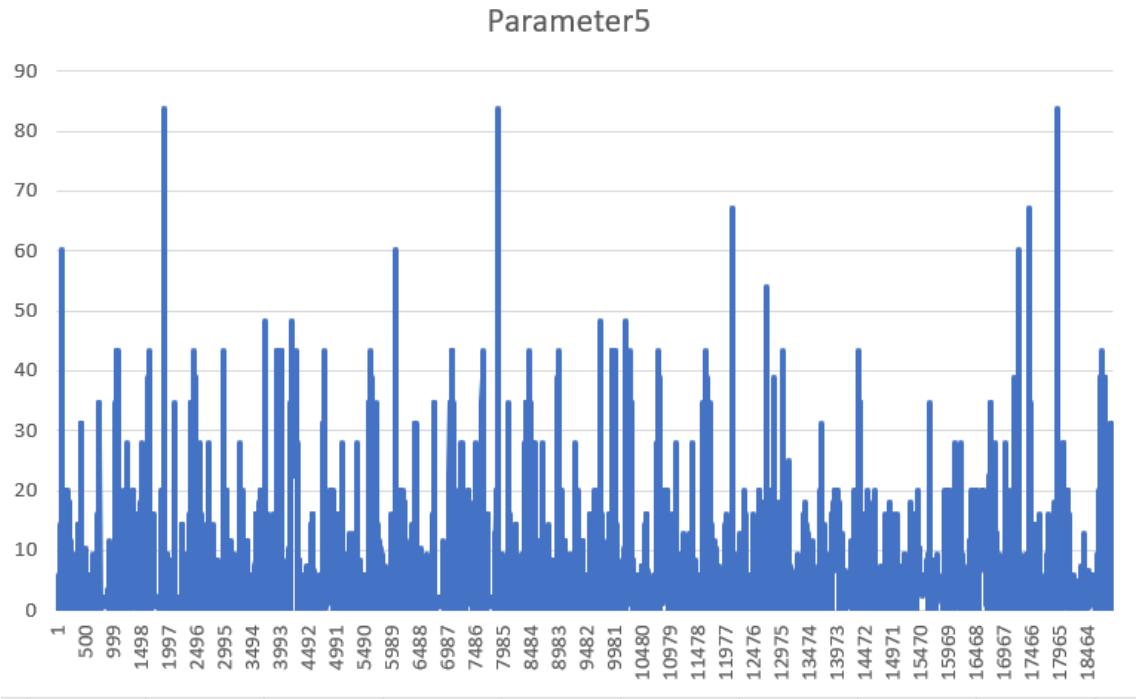

图 2

可见 Parameter1-4 大部分数据为长尾分布，数据之间的极差很大，这对后续的特征工程会产生影响。

Parameter1-label 和 Parameter5-label 的分布图如下：

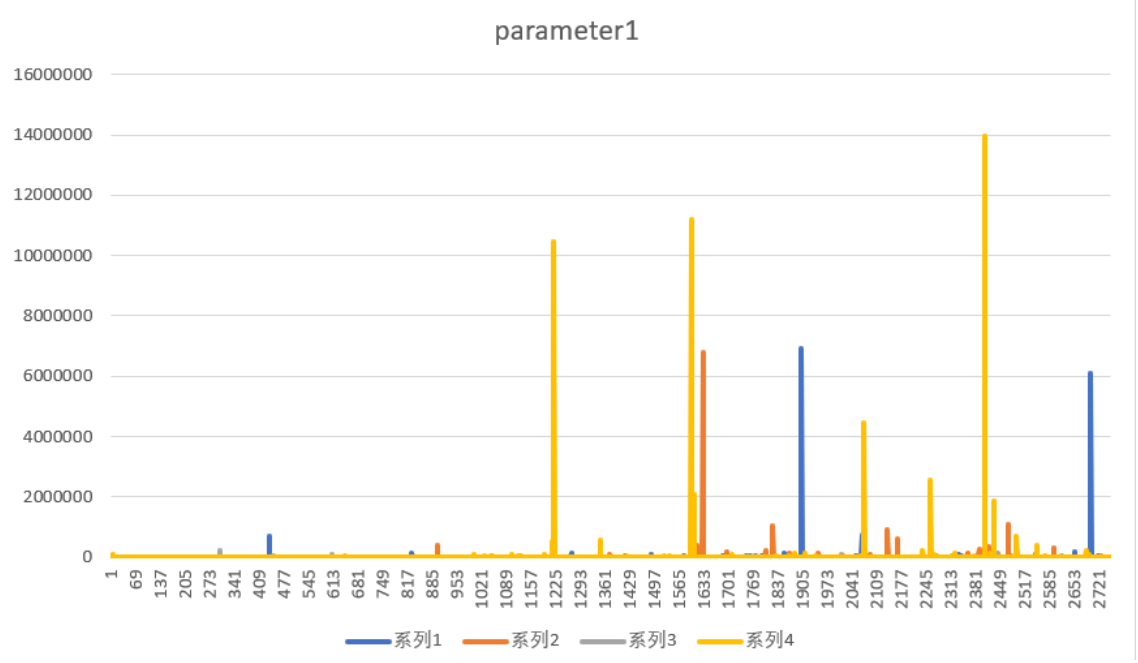

图 3

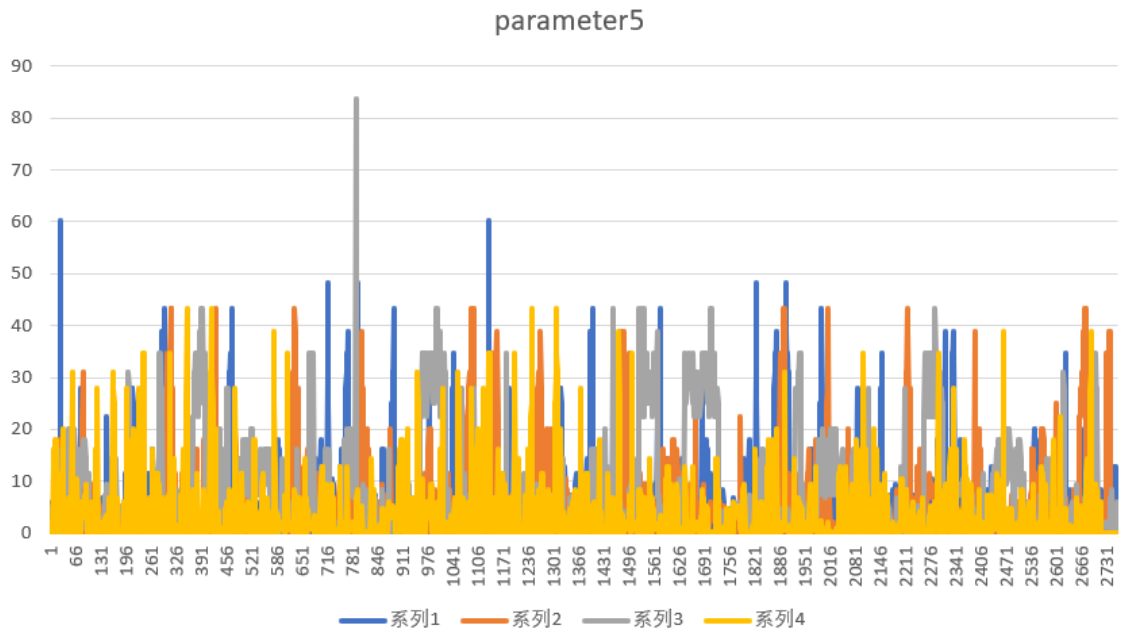

图 4

由上图可见，Parameter1 在四种质检指标上的分布基本一致，因此初步推断该特征对分类模型并不能提供有效的信息。相比之下 Parameter5 在四种质检指标上的分布各有不同，能够作为有效的输入特征。

在随机森林模型中，一开始我以十个工艺参数作为训练数据，训练后输出了 10 个参数的重要性指标，结果也验证了之前的分析，Parameter1-4 相对于 Parameter5-

10 在重要性上有一个数量级的差距。所以最终我选择了使用 parameter5-10 进行训练。

第 4 章 特征工程

数据无缺失值，无需补全，属性参数均为 Float 类型，质量指标共有四类，为字符串类型，数据均经过了脱敏处理。

质检指标 Quality_label 有四类：Excellent，Good，Pass，Fail。在不同的模型中我使用了不同的编码方式，在神经网络中我使用的是 one-hot 编码方式，利用 keras 中的 np_utils.to_categorical 函数实现将类别变量（categorical function）的输出标签转化为数值变量。

而在随机森林和梯度提升树中，我采用了 factorize 方法，利用 pandas 的 factorize 方法，将四个属性映射为 0-4 的数字，相对 one-hot 的编码方式更节省空间，这点在类型比较多时效果更为明显。

第 5 章 模型训练与调优

5.1 随机森林

随机森林和梯度提升树都是基于决策树的模型，但是相对后者，随机森林采用了 bagging 集成学习，更不易发生过拟合，模型应用范围更大。调优部分调用了 sklearn 库的 GridSearchCV 方法进行网格搜索，对参数 n_estimators 迭代次数，max_depth 决策树最大深度，min_samples_split 节点可分的最小样本数，min_samples_leaf 叶子节点含有的最少样本，max_features 构建决策树最优模型时考虑的最大特征数，分步进行调优以减少运算时间，其中 min_samples_leaf 和 min_samples_split 相互依赖只能一起调优。

5.2 梯度提升树

采用了 Catboost 模型（见第二章介绍），决策提升树本质是集成学习 boosting + 决策树。调优部分对学习率和决策树个数等参数进行了网格搜索，同样调用了 sklearn 库的 GridSearchCV 方法进行网格搜索。

5.3 神经网络

神经网络采用了 DNN 模型，尝试了多种神经网络模型，最终设计的神经网络

结构如图 5 所示，由 20 层的输入层，层数为 40 层的隐藏层，一层输出组成，隐藏层激活函数全部使用 relu 函数，输出层激活函数采用 softmax。调优部分对 optimizers，epochs 等进行了网格搜索。

尝试了很多神经网络模型结构，发现结构复杂的模型未必分类效果更好，但是训练时间却增加很多，最终还是选择了这个简单的模型，运行相对较快而且准确率较高。

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 20)                140
_________________________________________________________________
dense_2 (Dense)              (None, 40)                840
_________________________________________________________________
dropout_1 (Dropout)          (None, 40)                0
_________________________________________________________________
dense_3 (Dense)              (None, 4)                 164
=================================================================
Total params: 1,144
Trainable params: 1,144
Non-trainable params: 0
```

图 5

第 6 章 模型融合

模型融合可以采用线性加权融合，为每个模型输出的结果分配一个系数，然后相加求和，减少错误的样本带来的误差。（由于网站原因，输出 submission 文件已无法进行线上评测，而模型融合实际上是对多种模型得到的 submission 文件进行加权平均，因而实际上模型融合已没有意义）

第 7 章实验

7.1 实验设计

7.1.1 随机森林

随机森林调用了 sklearn 库的 GridSearchCV 方法进行网格搜索，分别对 n_estimators，max_depth，min_samples_split，min_samples_leaf，max_features 分步

网格搜索调优，其中 min_samples_leaf 和 min_samples_split 相互依赖只能一起调优。采用准确率作为评价指标（赛题中使用 MAE 系数作为评价指标，但是由于网站原因目前并不适用）。

7.1.2 梯度提升树

梯度提升树同样进行网格搜索调优，对学习率和决策树个数等参数进行了网格搜索。采用准确率作为评价指标。

7.1.3 神经网络

训练不同的神经网络网络模型并对比；对 optimizers，epochs 等进行网格搜索调优。采用准确率作为评价指标。

7.2 对比实验

采用十折交叉验证，分别对比调优后各模型的准确率和各个模型的准确率。

7.3 实验结果

四分类问题，三种方法的准确率经多次实验，多分类准确率都在 50%-60%之间

模型	多分类准确率	训练用时
随机森林	0.56	0.47s
梯度提升树	0.57	6.01s
神经网络	0.54	约 10 分钟

第 8 章结论

1. 树模型在 multi-class 问题上表现良好，随机森林和梯度提升树都表现出了良好的分类准确率
2. DNN 神经网络取得了和树模型相近的分类准确率，但是相比于树模型训练用时更长。神经网络在 multi-calss 问题上需要更适合的模型
3. 三种模型对于 multi-class 问题都表现出了相近的准确率，模型融合后可以进一步提高分类准确率
4. 从实验结果来看，树模型多分类准确率和 DNN 神经网络接近，训练时间上树模型远短于神经网络，相比之下树模型更适合于该赛题

参考文献

[1] 爱丽丝•郑，阿达曼・卡萨丽。 《精通特征工程》

[2] Keras 官方文档 https://keras.io/zh/

[3] 锡南・厄兹代米尔，迪夫娅・苏萨拉 《特征工程入门与实践》

[4] 陈旭梅，龚辉波，王景楠 基于 SVM 和 kalman 滤波的 BRT 行程时间预测模型研究[J]. 交通运输系统工程与信息