Modeling and Simulation 建模与仿真, 2022, 11(3), 734-743
Published Online May 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.113069

Hans汉斯

基于XGBoost的双目标零件类别划分技术与优化

朱　露，仲梁维

上海理工大学机械工程学院，上海

收稿日期：2022年3月18日；录用日期：2022年5月14日；发布日期：2022年5月20日

摘　要

针对目前复杂产品装配对象内部零件结构层次识别困难的情形，以提高装配工艺文件编制的合理性、准确性以及其效率为目标。研究基于贝叶斯优化的XGBoost算法，以五套不同型号的生物器皿消毒机三维模型与装配工艺卡作为工艺数据源，对零件进行了关于结构类型和功能类型的双目标多类别的自动识别。得到结构类型和功能类型识别的准确率分别可达到96%和90%。实现自动识别并划分零件类型，丰富了装配体内部信息。

关键词

装配工艺，零件划分，XGBoost，贝叶斯优化算法

Classification Technology and Optimization of Double Objective Parts Based on XGBoost

Lu Zhu, Liangwei Zhong

School of Mechanical Engineering, Shanghai University of Science and Technology, Shanghai

Received: Mar. 18th, 2022; accepted: May 14th, 2022; published: May 20th, 2022

Abstract

In view of the difficulty in identifying the internal part structure hierarchy of complex product assembly object, the goal is to improve the rationality, accuracy and efficiency of assembly process documentation. The XGBoost algorithm based on Bayesian optimization algorithm is studied. Five sets of three-dimensional models and assembly process cards of different types of biological utensils disinfection machines are used as the process data source to automatically identify the struc-

文章引用: 朱露, 仲梁维. 基于 XGBoost 的双目标零件类别划分技术与优化[J]. 建模与仿真, 2022, 11(3): 734-743.
DOI: 10.12677/mos.2022.113069

ture type and function type of parts. The recognition accuracy of structure type and function type can reach 96% and 90% respectively. Automatic identification and classification of parts are realized, which enriches the internal information of assembly.

Keywords

Assembly Process, Part Classification, XGBoost, Bayesian Optimization Algorithm

1. 引言

对于一个结构复杂，装配困难的装配对象而言，内部的零件也是多种多样的，而在对零件对象的信息进行自动提取的过程中，在没有人为对零件体类别进行划分的前提情况下，无法直接确定零件类别。装配对象内部的结构层次对于实例推理有着十分重要的影响，若无法对内部的零件进行类别上的分层，就无法清晰地解析装配对象内部的信息，从而影响推理的精确度，进一步阻碍较高层次的知识模式被发现。

在针对零件识别与分类方面文献[1]利用 Canny 边缘算子识别零件的图形轮廓并进行特征匹配以完成对工业零件的分类；文献[2]提出一种基于机器视觉的零件外形轮廓分类方法；文献[3]为了解决零件类型的确定问题，将卷积神经网络组织与数据分类进行匹配；文献[4]研究了图像特征提取的常见方法，并结合零件的特点，重点研究了零件的形状特征和几何特征提取方法。目前在对零件分类的研究上主要是采用图像识别方法[5] [6] [7]，本文创新性提出通过对零件的属性参数和特征参数的提取，利用 XGBoost 算法识别零件并按照结构类型和功能类型对其进行分类。

2. 零件类别概念分层的方案设计

2.1. 类别划分

装配对象实例化需要对零件这一对象进行合理的概念分层，零件类别属于标称属性，Solid Works 二次开发技术可以快速地获取零件的名称，却无法对未作标注的零件进行类别的提取，因此，概念分层的父层类别需要人工来定义。在进行概念分层之前，需要明确分层的结构以及对零件父层类别进行自动划分。如图 1 所示。

零件大类由其功能与结构划分人为定义主要为以下几种：

1) 按功能划分：连接件(A)，紧固件(B)，定位件(C)，密封件(D)，传动件(E)，支承件(F)；

2) 按结构划分：轴类件(G)，盘类件(H)，肋板件(I)，箱盖件(J)，叉架件(K)；

3) 其他标注：其他件(M)。

在以往人工对零件进行编码时其规则主要通过“功能编码－结构编码”的方法，例如对于法兰盘这一零件，它属于盘类零件，主要是起到轴向定位的作用，因此它的编码为“C-H”。如果无法对某零件的结构或者功能有清晰地了解，可以通过其他件(M)进行编码，例如对于花键而言，它是起到了连接的作用，而结构却无法进行清晰地辨认，因此它的编码可以是“A-M”。利用这种简单的编码方法对零件大类进行划分，零件有了清晰的归属，也使得装配对象有了更加丰富的内在信息[8]，但在数据量巨大的情况下

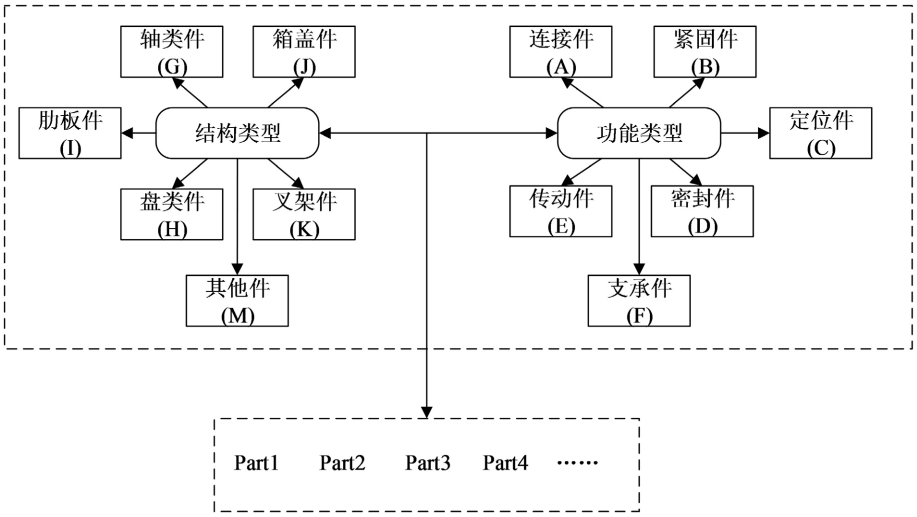


Figure 1. Conceptual hierarchical design based on part category
图 1. 基于零件类别的概念分层设计

会导致时间成本上升。实际上对于零件体对象而言，其本身存储了一系列可以代表该零件整体参数的信息，即零件信息库，该库主要是由零件的自然属性，轮廓属性，特征属性，以及零件基本类型组合而成，XGBoost 可以拟合零件特征参数以达到对零件类型自动标记的目的，这本质上是一个多分类问题。

2.2. XGBoost 方法

XGBoost 算法的是每棵回归树、每个叶子节点都会有自己的得分[9]，得分相加就是最终预测结果的产生，为了学习得到模型中的最优参数，对于分类问题，可以采用逻辑回归函数映射成概率，以下就是最小化的正则化目标[10]：

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{K=1}^{K} \Omega\left(f_k\right) \tag{1}$$

$$\Omega\left(f_k\right) = \Upsilon T + \frac{1}{2}\lambda\|w\|^2 \tag{2}$$

这里的 $l\left(y_i, \hat{y}_i\right)$ 是一个可微的损失凸函数，而 $\Omega\left(f_k\right)$ 则是惩罚函数，它可以额外地增加正则化项，这有助于平滑最终学习的权值，以避免过拟合。模型的预测精度主要是由方差与偏差决定，损失函数代表了偏差；正则化的目标将倾向于选择使用简单和预测函数的模型，更简单的模型就代表了更小的方差。在整个正则化函数中 T 代表了叶子节点的个数， w 代表叶子节点的分数， Υ 和 λ 是正则化系数，它们分别控制树的叶节点数量和节点值的总体大小。

3. 双目标划分

由于零件的类型由结构与功能两大类组成，如果直接让 XGBoost 预测结构与功能的组合，目标值将有数十种，这会导致分类较多，每一类的样本量在采集初期较少，没有足够的样本作为训练集去拟合模型，因此本文分两大类进行双类别进行划分，大大减少了划分类型的个数，相对增加了样本量，保证每一类有足够的样本进行训练，在最后进行组合，设计具体的方案流程如图 2 所示。

零件信息库中包含了半自动化采集到的所有零件样本，在训练 XGBoost 模型之前，需要将样本从数据库中提取出来，同时进行数据预处理，由于零件的材料、类型等属于离散化属性，即字符串类型的数据，因此需要进行独热(One-hot)编码进行多列扩展，延展成稀疏矩阵，如图 3 所示为部分数据的稀疏矩阵。

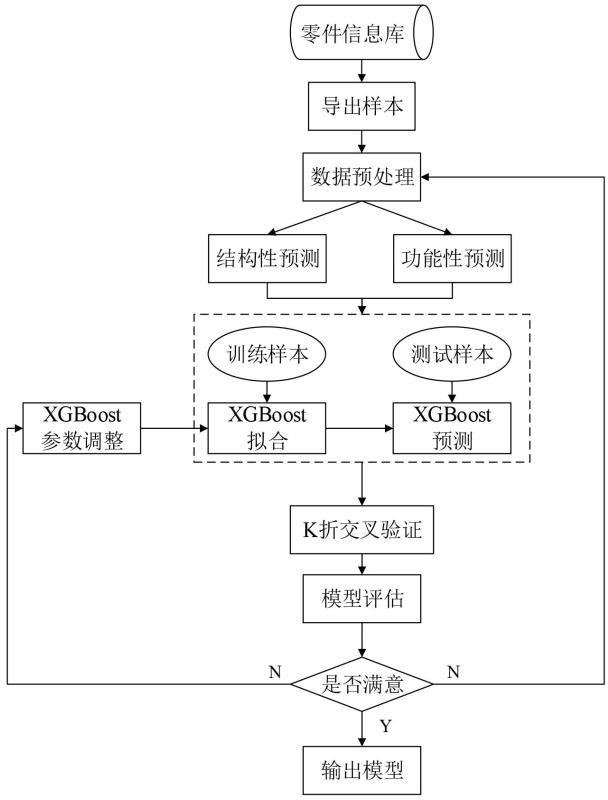


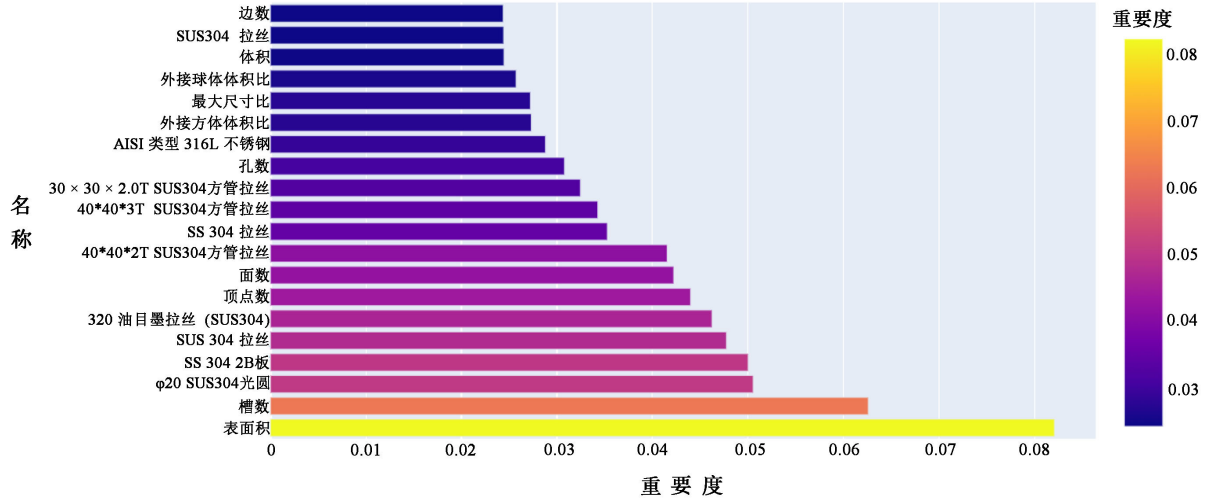
Figure 2. Model training flow chart of XGBoost
图 2. XGBoost 的模型训练流程图

	0Cr18Ni10(GB)	30*30*2.0T SUS304方管拉丝	30*30*2T SUS304 方管拉丝	30×30×2.0T SUS304方管拉丝	320 油目墨拉丝 (SUS304)	320 目油墨拉丝 (SUS304)	320 目油墨拉丝 (SUS304)	320 目油磨拉丝 (SUS304)	320 目油墨拉丝 SUS 304	40*40*2T SUS304 方管拉丝	…	硅	硅橡胶	硅胶	硅胶垫	胶木	铍红铜，UNS C17000	锌合金 7; AG40B; Zn-4Al-0.015Mg	镀蓝白锌板
	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0

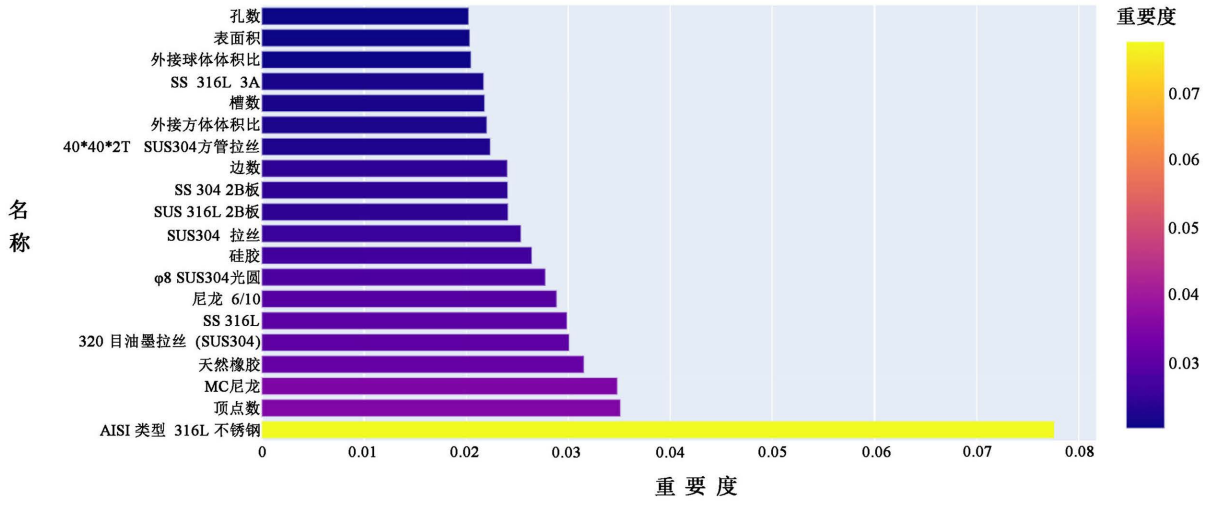
Figure 3. Material sparsity matrix of parts
图 3. 零件的材料稀疏矩阵

在完成上述数据处理之后，利用 XGBoost 初始化状态模型(超参数默认状态)对给定的零件特征重要性进行排序，观察对模型影响最大的前 n 个特征，其主要的评估依据包括有：1) 对生成叶子节点时特征所产生的增益大小；2) 在所有树种该特征所代表的子节点个数与总个数的百分比，即覆盖度；3) 特征在模型树种发生的相对次数，即频率。目前零件信息库中，零件特征的重要性排序按照功能和结构划分前 20 重要的特征如图 4 所示。

在影响零件类型最为重要的前 20 个的特征中，绝大部分还是材料属性。除此之外结构类别的划分更加依赖于零部件的特征属性，例如：表面积，槽数，顶点数，面数。而功能类别的划分更加依赖于零件的材料，其中 316L 不锈钢成为材料属性中对分类器贡献度最高的特征。



(a)



(b)

Figure 4. XGBoost-based part feature correlation ranking. (a) Importance of features related to structure type; (b) Importance of features related to functional types

图 4. 基于 XGBoost 的零件特征相关度排序。(a) 与结构类型相关的特征重要性；(b) 与功能类型相关的特征重要性

为提高模型的预测精度，需要对 XGBoost 的超参数[11]做初始化设置，主要参数名与其对应的参数选择值如表 1 所示：

将重新设置参数的 XGBoost 模型用来拟合数据预处理结束后的训练集，选取由 100 个基分类器组成的集成分类器的第三棵树进行观察，如下图 5 所示。

可以观察到树的生长状态，首先从 SS316L 材料属性作为根属性进行延伸，第二层的划分特征为顶点数，第三层是槽数，第四层为面数与体积，第五层输出子叶节点，保证每次延伸的增益大于 0.1，且保证树的深度小于 5 层，防止树结构太复杂而导致的模型过拟合。

Table 1. Selection of initial hyperparametric values of XGBoost
表 1. XGBoost 初始超参数值选择

参数名	参数值	说明
n_estimators	100	基分类器个数
booster	gbtree	指定基分类器
objective	softmax	指定学习任务和相应的学习目标或使用的自定义目标函数
gamma	0.1	损失下降多少才进行分裂，控制叶子节点的个数
max_depth	5	树的深度，越大越容易过拟合
reg_lambda	3	控制模型复杂度的权重值的 L2 正则化项参数
subsample	0.7	随机采样训练样本
eta	0.1	学习率
nthread	4	线程数
min_child_weight	1	子节点中最小的样本权重和

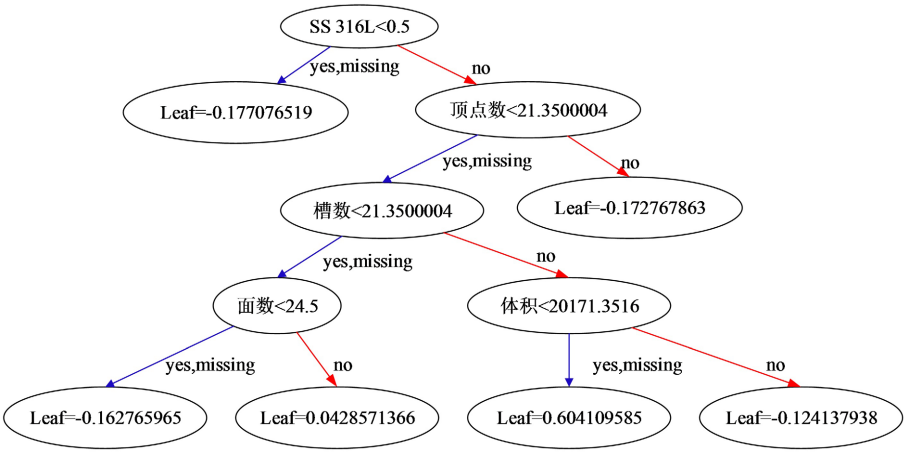


Figure 5. The third tree model structure of XGBoost model of structure type
图 5. 结构类型 XGBoost 模型的第三棵树模型结构

K 折交叉验证就是将零件样本分解为 K 份，循环 K 次进行模型的拟合，其中 K − 1 份为训练集，1 份为验证集，利用每轮循环中的得分求算术平均数，这样的得分可以更好的得出模型的泛化性能。根据 10 折交叉验证所得模型的得分评估报告如下表 2 和表 3 所示。

Table 2. XGBoost model evaluation (structure type)
表 2. XGBoost 模型评估(结构类型)

10 折交叉验证平均准确率：		0.94	
类型	平均查准率	平均查全率	平均 F1 值
盘类	1.00	0.5	0.67
箱盖类	0.91	0.94	0.92
肋板件	0.92	0.95	0.94
轴类	1.00	0.71	0.83
其他件	0.97	0.97	0.97
平均	0.96	0.81	0.88

Table 3. XGBoost model evaluation (function type)
表 3. XGBoost 模型评估(功能类型)

10 折交叉验证平均准确率：		0.88	
类型名称	平均查准率	平均查全率	平均 F1 值
传动件	1.00	0.5	0.67
其他件	0.91	0.94	0.92
定位件	0.92	0.95	0.94
密封件	1.00	0.71	0.83
支承件	0.97	0.97	0.97
紧固件	1.00	0.82	0.90
连接件	0.80	0.91	0.85
平均	0.94	0.82	0.87

其中查准率表示的意思是预测是正例的样本中被正确分类的样本的比例。查全率表示的意思是真实是正例的样本中被分类正确的样本的比例。F1 值是对查准率与查全率得一种综合的衡量标准，F1 值越接近 1，模型的训练效果越好。由上述的两个模型评估报告可以看出模型的性能不错，但仍然有很大的提升空间，尤其是对功能类别的划分，由于样本的标签分布很多时候是不平衡的，因此在一些类型在分类器上的表现跳跃性较强，继续采集稀缺样本，以及升采样与降采样的方法可以解决样本不平衡问题。

4. 优化

模型的效果提升包括增加特征，筛选异常特征以及调整模型的超参。在假定数据集的数据质量无法再进行优化之后，模型的效果提升主要依靠超参数的调节[12]，这里主要是采用贝叶斯优化方法(BOA)来寻找最优参数[13]，这种方法极大地减少了复杂模型的迭代时间。BOA 模型通过先验模型(prior function)来描述目标点的分布[14]，利用 AC 函数来选取新的点，这是模型的两大核心组件，点的质量决定了 BOA 的收敛速度与最终解的质量，如果选点不当，很容易形成进入局部最优解的情况，因此需要多方面的权衡，然后综合选点。

贝叶斯优化算法需要关注的是模型的输入参数，参数的取值范围，迭代次数以及输出参数，对于输入参数，只需要给定每个参数的范围，算法会自动从范围中选定初始化点位，逐步向最优逼近，在对双目标零件类型进行自动划分之前，对模型的超参数已经做了初始化设置。这里主要调节的参数以及设置的取值范围如表 4 所示。

Table 4. Bayesian optimization of XGBoost hyperparameters
表 4. 贝叶斯优化 XGBoost 超参数

参数名	说明	取值范围
n_estimators	基分类器个数	[50~150]
gamma	损失下降多少才进行分裂，控制叶子节点的个数	[0.05~0.1]
max_depth	树的深度，越大越容易过拟合	[3~15]
subsample	随机采样训练样本	[0.6~0.9]
eta	学习率	[0.01~0.1]
min_child_weight	子节点中最小的样本权重和	[1~7]

在设置完调节参数之后，还需要对迭代次数进行设置，贝叶斯迭代的方法主要分为两种[15]，一种是探索，就是点的选择位置会尽可能远离已知点，点的分布会尽可能平均，而另一种则是利用，即尽量挖掘已知点周围的点，使其进入一个局部最大的情况。这里选择探索的次数为 5 次，利用的次数为 25 次，进行总和为 30 次的优化。

优化之后仍然采用 10 折交叉验证的方法检验模型的泛化性能，将最终的平均准确率得分作为输出结果。调优结束之后，发现零件结构类型和功能类型的分类任务的最终得到最优参数组合为如表 5 所示。

Table 5. Hyperparameters adjustment after Bayesian optimization
表 5. 贝叶斯优化后超参数调节情况

参数名	结构分类	功能分类
n_estimators	79	97
gamma	0.07	0.06
max_depth	14	12
subsample	0.63	0.62
eta	0.08	0.09
min_child_weight	1.04	1.41

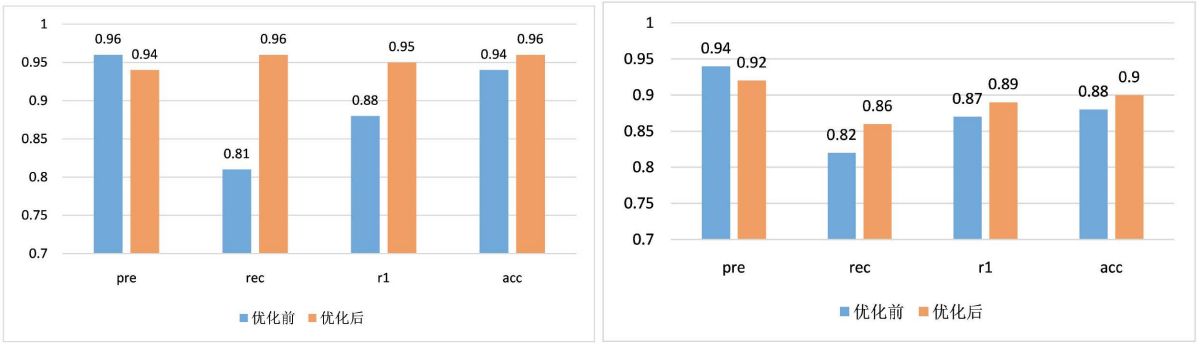
将优化之后的模型超参再次带入模型之中，进行 10 折交叉验证法，可以得到优化后的模型评估报告如表 6 和表 7 所示，与优化前的模型效果对比如图 6 所示。

Table 6. Evaluation report of optimized XGBoost model (structure type)
表 6. 优化后的 XGBoost 模型评估报告(结构类型)

10 折交叉验证平均准确率：		0.96	
类型名称	平均查准率	平均查全率	平均 F1 值
盘类	1.00	0.67	0.80
箱盖类	0.87	0.96	0.92
肋板件	0.93	0.92	0.93
轴类	1.00	1.00	1.00
其他件	0.97	0.97	0.97
平均	0.94	0.96	0.95

Table 7. Evaluation report of optimized XGBoost model (function type)
表 7. 优化后的 XGBoost 模型评估报告(功能类型)

10 折交叉验证平均准确率：		0.90	
类型名称	平均查准率	平均查全率	平均 F1 值
传动件	1.00	0.90	0.95
其他件	0.88	0.94	0.91
定位件	0.93	0.76	0.84
密封件	0.95	0.90	0.93
支承件	0.88	0.91	0.89
紧固件	0.92	0.80	0.86
连接件	0.93	0.85	0.89
平均	0.92	0.86	0.89



(a) (b)

Figure 6. Evaluation of model effect before and after Bayesian optimization. (a) Effect comparison of structural category model before and after optimization; (b) Effect comparison of function category model before and after optimization
图 6. 贝叶斯优化前后模型效果评估。(a) 结构类别模型优化前后效果对比；(b) 功能类别模型优化前后效果对比

通过贝叶斯优化后的模型在同一套数据集上明显优于优化前期，除了查准率有了些许下降之外，其他指标都有一定的提升，其中最明显的是结构类型划分模型的查全率，有了 15%的提升，准确率也分别提升了将近 2%，优化效果较为明显。

此次选择清洗机前导轨验证上述算法对装配对象内部零件识别与分类的准确性，模型中包括的零件有导轨、支架、电机罩等，其三维模型如图 7 所示。

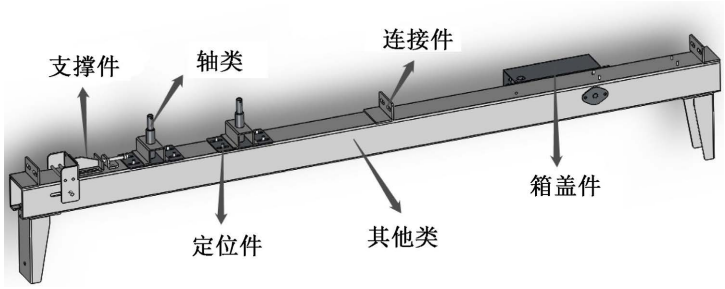


Figure 7. Three dimensional model of front guide rail of cleaning machine
图 7. 清洗机前导轨三维模型图

分类后前导轨组装子零件类型与如表 8 所示，利用该方法能较为准确地给装配体内部零件划分类别。

Table 8. Front rail assembly internal part type information
表 8. 前导轨组装内部零件类型信息

类型	轴类	盘类	肋板件	箱盖类	传动件	紧固件	密封件	连接件	定位件	支撑件	其他件
个数	7	0	2	1	2	0	2	5	16	8	35

5. 总结

本文主要任务是利用概念分层的思想，通过基于贝叶斯优化的 XGBoost 模型对零件信息库中的零件类型进行双目标分类，之后将零件的功能类型与结构类型进行组合编码，输出零件的基本大类。通过该方案，大大丰富了零件信息库中零件本身的价值，同时又重新构建了新的特征可用于检索与筛选。其更大的意义在于使得装配对象有了更加丰富的内部信息，在装配实例推理的过程中扮演了重要的角色。目

前，提取的零件信息主要来源于五套不同型号的生物器皿消毒机模型，总共 10,832 个样本，优化后的模型在结构类型的划分中可以达到 96%的准确率，在功能类型的划分中可以达到 90%的准确率。该结果仍然有提升空间，如何进一步提高本文方法的通用性与泛化能力，增强数据集包容性，是接下来的研究方向。

参考文献

[1] 王磊. 基于 canny 边缘检测的工业零件分类识别[J]. 电子设计工程, 2019, 27(21): 190-193.

[2] 王晓初, 邱杰豪, 欧阳祥波, 简川霞, 范彬祥. 基于机器视觉的轴承盖外形轮廓分类方法[J]. 包装工程, 2020, 41(23): 217-222.

[3] Pechenin, V.A., Bolotov, M.A. and Yu Pechenina, E. (2020) Neural Network Model of Machine Parts Classification by Optical Scanning Results. *Journal of Physics*: *Conference Series*, **1515**, Article ID: 052008. https://doi.org/10.1088/1742-6596/1515/5/052008

[4] 司小婷. 基于视觉的零件特征识别与分类方法研究与实现[D]: [硕士学位论文]. 沈阳: 中国科学院研究生院(沈阳计算技术研究所), 2016.

[5] Joshi, K.D., Chauhan, V. and Surgenor, B. (2020) A Flexible Machine Vision System for Small Part Inspection Based on a Hybrid SVM/ANN Approach. *Journal of Intelligent Manufacturing*, **31**, 103-125. https://doi.org/10.1007/s10845-018-1438-3

[6] Krüger, J., *et al*. (2019) Deep Learning for Part Identification Based on Inherent Features. *CIRP Annals—Manufacturing Technology*, **68**, 9-12. https://doi.org/10.1016/j.cirp.2019.04.095

[7] Feng, Y.Q. and Li, B. (2010) Solid Model Reconstruction and Feature Recognition for Mechanical Part Based on Slice Image. *Proceeding of International Conference on Computer Science & Education*, *ICCSE* 2010, Hefei, 24-27 August 2010, 221-224. https://doi.org/10.1109/ICCSE.2010.5593651

[8] 吴炜. 烟草机械产品零件分类编码系统研究与开发[D]: [硕士学位论文]. 上海: 上海交通大学, 2012.

[9] 许裕栗, 杨晶, 李柠, 甘中学. XGBoost 算法在区域用电预测中的应用[J]. 自动化仪表, 2018, 39(7): 1-5.

[10] 付宇. 基于XGBoost模型的贵州省野外真菌菌种预测系统及可视化展示[D]: [硕士学位论文]. 北京: 北京林业大学, 2020.

[11] 孙俊佚雄. 基于大数据的智能家电故障检测和诊断模型研究[D]: [硕士学位论文]. 桂林: 桂林电子科技大学, 2020.

[12] Awojoyogbe, O.B. (2003) A Mathematical Model of Bloch NMR Equations for Quantitative Analysis of Blood Flow in Blood Vessels of Changing Cross-Section-PART II. *Physica A Statistical Mechanics & Its Applications*, **323**, 534-550. https://doi.org/10.1016/S0378-4371(02)02025-3

[13] 李叶紫, 王振友, 周怡璐, 韩晓卓. 基于贝叶斯最优化的 XGBoost 算法的改进及应用[J]. 广东工业大学学报, 2018, 35(1): 23-28

[14] Zhang, B., Peng, M., Cheng, S., *et al*. (2020) A Decision-Making Method Based on Bayesian Optimization Algorithm for Small Modular Reactor. *Kerntechnik*, **85**, 109-121. https://doi.org/10.1515/kern-2020-850208

[15] 邓帅. 基于改进贝叶斯优化算法 CNN 超参数优化方法[J]. 计算机应用研究, 2019, 36(7): 1984-1987.