工业大数据分析案例剖析

工业互联网产业联盟（AII）

2021 年 12 月

工业互联网产业联盟

Alliance of Industrial Internet

编写团队

主编：郭朝晖、王晨、王建民

案例 1：韩晓明、高艳辉、李建国

石家庄开发区天远科技有限公司

案例 2：胡鹏程、谢杰君、陈吉红

华中科技大学机械科学与工程学院、国家数控系统工程技术研究中心

案例 3：郭朝晖

上海优也信息技术有限公司

案例 4：杨明明，刘贤康，李佳鹤

浙江蓝卓工业互联网信息技术有限公司

工业互联网产业联盟

Alliance of Industrial Internet

目　录

工业互联网产业联盟

Alliance of Industrial Internet

绪　论

工业大数据分析作为工业智能化发展的核心之一，是实践性非常强的工作，现实中的失败比例非常高。在《工业大数据分析指南》中虽然已对通用的工业大数据分析方法和分析流程进行了归纳和总结，但其更加关注于具有普遍指导意义的方法论，为能更好的指导企业开展工业大数据分析实践，我们选取了四个不同行业中已经落地应用的典型案例，并依照《工业大数据分析指南》的方法体系进行了较为深度的剖析，形成了本案例集。

人们用工业大数据分析的办法来发现知识并指导行动。如果错误的认识误导了行动，可能会给工业企业带来非常严重的后果。所以，工业界在实际应用中，对数据分析结果的可靠性要求很高，这对于工业大数据分析应用的落地带来了极大的挑战。

传统概率统计方法是从基本的理论假设开始展开研究，分析结果的可靠性是由理论前提和假设保证的。科学家从事科研工作时，可以根据分析工作的需求去采集和配置数据，从而得到可靠的结果。工业大数据分析则是通过数据本身表现出来的特点来发现规律。从事工业大数据分析时，人们往往只能根据工业现场已有的数据进行分析，而这些数据往往不是为特定数据分析工作而准备的。在某些场景下，数据可能从根本上就无法支撑分析的目标。从这种意义上说，特别是数据量无法达到一定规模的情况下，基于统计方法的数据分析建模不能够准确捕捉到工业现场问题与征兆之间的因果性，因而单纯依靠数据分析模型做出决策存在

相当程度的不确定性。

在追求确定性的过程中，有两种常见的挑战：一种是模型混淆了相关与因果，一种是把特殊条件下的因果关系扩大到一般情况。要应对这两种挑战，人们都必须借助对工业机理的认识。实践证明：如果仅仅以精度为标准衡量模型和结果的好坏，就很难保证成果的可用性，必须善于利用工业机理来选择数据、分析结论。从这种意义上说，工业大数据的分析在实际应用的落地过程，也是工业机理与数据分析的融合过程。

本案例集选取的四个典型案例各有特色，案例 1 是工程机械行业中利用机器学习技术开展的故障诊断实践，首先对液压泵的故障形成机理进行了完善的分析，将问题转换为预测泵压，并建立了多个变量与泵压之间的关联分析，然后建立机器学习模型，这是一种典型的利用领域知识完成问题建模和特征抽取，然后辅以数据分析方法建模的思路。

工艺参数优化是工业大数据分析一类特别典型的问题。案例 2 选择在数控加工领域，关注效率、质量与加工成本的多目标优化问题，如何通过工艺参数有效地实现多个目标的同时优化。方案一针对粗加工场景关注单次切削时间而非质量，通过试切操作收集一系列数据，建立了单目标优化求解模型。方案二以日常加工任务数据生成样本数据，以进给速度、主轴转速、切宽及切深建立能够表征加工过程物理状态的主轴功率（代替铣削力）预测模型，进而针对该模型进行工艺参数优化。在这个案例中我们既可以看到如何直接利用优化模型进行较为粗颗粒度的控制调整，

也可以参考对控制过程建立拟合的机器学习模型，进而实现对过程感知的白盒调整方法。

案例 3 是面向材料微观结构性能规律的研究，是较为少见的探索如何在研发过程利用数据方法建模的例子。案例 3 针对钢铁材料的强度、塑性、韧性、硬度等力学性能预测问题，通过数据分析方法，既减少标准试验的代价，也试图建立面向研发过程的相对准确、可靠的规律性结论。钢铁晶体组织的形成是一个动态的过程，由于大量研究结果是在特定成分和工艺条件下数据拟合的结果，没有统一的理论公式，因而，如何通过数据方法实现对规律性知识的总结是这个案例带给我们的重要启示。

案例 4 针对橡胶行业中胶粒水溶液的浓度测量问题，基于工艺流程设计、生产过程数据、设备运行机理等多维信息，建立胶粒水溶液的软测量方法预测模型，进而对卤化反应阶段的胶浆浓度进行有效控制，提升装置的综合运行效能。这是流程制造业中一类非常典型的问题，大型罐体内部非常难以实现在线测量，当前多依靠人工经验进行控制，如何充分利用操作经验、运行机理与数据科学融合的建模方法，实现对目标物理量的软测量方法，仍是业内的难题之一。本案例另一个非常有启发的点是区分了稳定和非稳定工况，分别采取了深度学习预测与机器视觉识别的方法，这种复合式方法对于实际落地应用效果会有较大的提升作用。

在本文的四个各具特点案例中，我们可以看到工业大数据的实践者们，如何将行业知识、机理与数据分析方法结合起来，通过从业务理解一直到模型落地的闭环过程解决实际业务问题。我

们希望通过对这些案例的深度剖析，把其中的成功经验分享给大家，帮助读者少走弯路并带动这一领域的技术发展。目前，人们对工业大数据分析的认识还需要不断深入，但我们相信，随着数据技术的不断发展，数据条件会越来越好，成功的应用也会越来越多。

一、案例 1：挖掘机液压泵故障预测—天远科技

（一）案例背景

挖掘机又称挖掘机械，是用铲斗挖掘高于或低于承机面的物料，并装入运输车辆或卸至堆料场的土方机械。挖掘机挖掘的物料主要是土壤、煤、泥沙以及经过预松后的土壤和岩石。从近几年工程机械的发展来看，挖掘机的发展相对较快，挖掘机已经成为工程建设中最主要的工程机械之一。由于挖掘机常常处于恶劣的工作环境下，故障率持续上升，一旦出现重大故障，造成停机，轻则造成延误工期等经济损失，重则危害车上人员的生命安全。因此通过机器学习手段提前预测挖掘机零部件故障具有至关重要的意义。

发动机，液压泵，分配阀是人们常说的挖掘机三大件，挖掘机不像汽车一样由发动机提供动力，经过变速箱、传动轴驱动整车前进，而是通过发动机带动液压泵转动，由高压液压油通过液压马达、液压油缸等液压执行元件带动整车动作。

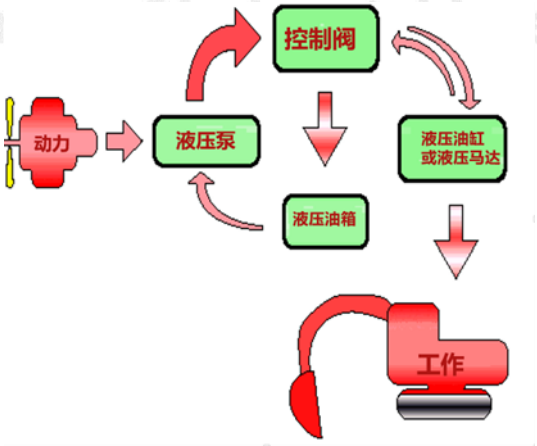


图 1 挖掘机工作示意图

液压泵是为液压传动提供加压液体的一种液压元件，是泵的一种。它的功能是把动力机（如电动机和内燃机等）的机械能转换成液体的压力能。影响液压泵的使用寿命因素很多，除了泵自身设计、制造因素外和一些与泵使用相关元件（如联轴器、滤油器等）的选用、试车运行过程中的操作等也有关系。

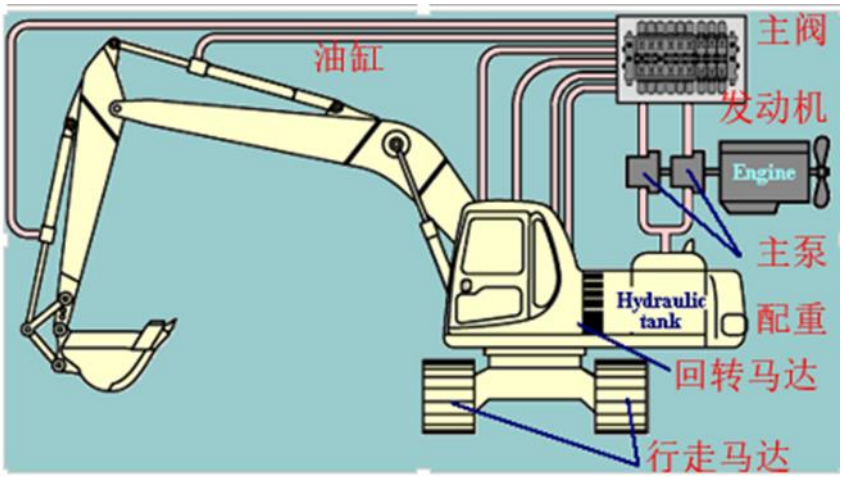


图 2 液压泵系统组成元件

液压泵是液压挖掘机中发生故障最多的元件，而液压泵一旦发生故障就会立即影响挖掘机液压系统的正常工作，甚至不能工作。液压泵对于挖掘机的重要性不言而喻，因此预测挖掘机液压泵的故障也是一个相当重要的课题。

液压泵主要有叶片泵、齿轮泵、柱塞泵三种类型，其常见故障主要包括:

第一，齿轮泵的常见故障大部分是由其内部摩擦副的磨损引起。其正常磨损使径向间隙和轴向间隙(即断面间隙)增大，齿轮泵内泄漏现象加重，严重时泵体内孔或两侧板无法修复。此外，轴的密封也是经常损坏的部件。

第二，叶片泵的正常磨损量很小，零部件使用寿命较长。造成叶片泵故障的主要原因是油液污染，这是因为叶片泵的运动副配合较精密，当污染物进入摩擦副后，容易产生异常卡滞或磨损。另外，叶片泵的自吸性能不如齿轮泵，特别是小排量的叶片泵更是如此，所以油液是否清洁和吸油是否畅通，是叶片泵运行中需要特别注意的两个问题。

第三，柱塞泵中的径向柱塞泵在结构和运动性能上的弱点是径向力较大、自吸能力较差以及柱塞与柱塞孔的配合精度高；轴向柱塞泵的零件加工精度要求高。所以柱塞泵对油液的清洁度要求高，亦即柱塞泵对油液的过滤精度要求比齿轮泵的高。

目前针对液压泵的故障维修，采用的多是事后维修。与之不同的是，预测检修可通过对液压泵之前的状态进行故障预测来安排检修活动，具有自动化、高效率等显著优势。液压泵的故障预测利用传感器来采集挖掘机的数据信息，借助合适的算法来评估液压泵的健康状态，在故障发生前对其进行预测。

（二）解决方案

1、业务理解

（1）认识工业对象

液压泵常见的故障原因主要可归纳为油品质低或油污染程度高、零件磨损两方面。

液压泵常见的故障表现有以下几点：

1）液压泵磨损严重，液压泵的转动会不均衡，产生异响；

2）液压泵磨损，内泄量增大，液压泵的出油量会减少，流

量低到一定程度时会导致压力低（流量低会导致压力低，但是不是唯一的原因）；

3）液压泵磨损后，壳体的泄漏量会增大，因为壳体泄漏的液压油直接返回油箱，没有经过散热器的散热，所以可能导致液压油温高。

（2）理解数据分析需求

数据分析需求：判定挖掘机液压泵是否故障。

需求理解：液压泵故障常与液压泵磨损或油液污染有关，且它的动力来源于发动机，因此特征选取方面需选取发动机的参数以及油液的相关参数。若液压泵产生故障，根据故障的三个表现，结合目前传感器的数据，可采用预测泵压的方式进行分析：若液压泵故障，则会导致产生的泵压（P）会与正常情况时的泵压（P'）产生偏差（$\triangle P$），若当天偏离度（$\triangle P/P$）超过某个值（w）的占比超过某个阈值（*threshold*）时，则判定为液压泵故障。

假设每天的泵压为 y_{real}，对应的预测值为 y，即当：

$$\frac{\sum_{i=0}^{N}(\frac{y_i - y_{real_i}}{y_{real_i}} > w)}{N} > threshold$$

其中：

N 为当天采集的数据量

w 为高偏离度的设定值

threshold 为设定阈值

则当计算结果大于设定阈值时，预测为故障；否则，预测为正常。

（3）数据分析目标及评估

数据分析目标:

1）根据传感器每天传入数据，选择合适的特征及模型，计算得到预测泵压值;

2）根据预测值和实际值，计算得到每条数据偏离度;

3）计算每天不同偏离度的占比;

4）根据不同偏离度占比随时间的变化图结合故障信息，得到 *w* 和 *threshold*;

5）将所得到的模型和阈值保存;

6）根据每日数据，得到当日故障预测结果，并将预测为故障的结果保存。

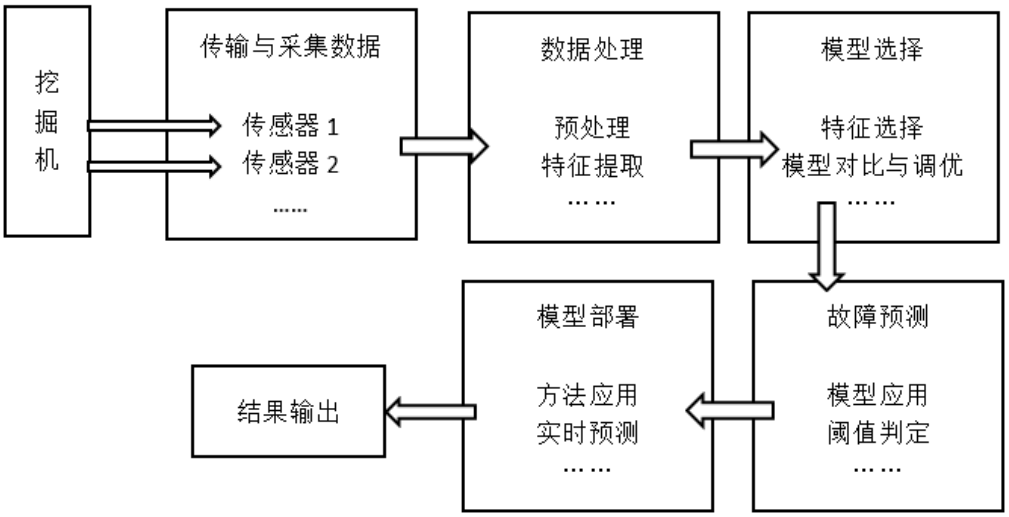


图 3 数据分析目标架构图

评估：预测准确度。最终预测结果的混淆矩阵查准率和查全率。

2、数据理解

（1）数据来源

传感器是一种检测装置，能感受到被测量的信息，并能将感受到的信息，按一定规律变换成为电信号或其他所需形式的信息输出，以满足信息的传输、处理、存储、显示、记录和控制等要求。

本案例采用的数据均来源于某台挖掘机传感器从 2020 年 6 月 1 日至 2020 年 10 月 7 日传入的每分钟实时数据，共计 101359 行数据，有 47 列参数值。

（2）数据分类及相互关系

所采集数据按数据取值范围分为离散型变量和连续型变量。连续型变量的值代表数值含义，但是离散型变量的值虽然也可能是数值型，但是并没有数值意义，需要经过处理后使用，常见的处理方式是将其转变成哑变量。

表 1 数据集主要变量描述

变量名	描述	类型
转速	发动机每分钟回转数	数值（0-2500）
扭矩百分比	实际发动机输出扭矩与发动机最大输出扭矩的比值	数值（0-100）
泵压	液压泵提供的压力	数值
模式	当前发动机工作模式	字符串
共轨燃油压力	燃油压力	数值
扭矩	当前实际扭矩值	数值
动作	挖掘机动作编码	数值
动作类型	动作类型	二进制字符串

（3）数据质量

1）完整性：采用数据采集率热力图来查看数据采集的完整

性。

2）规范性：查看每个字段数据类型及取值范围是否合理，不合理的取值当作异常值。

3）一致性：检查每个字段的数据类型。数据读入后，对应的每个字段的数据类型也会有变化，需要调整成正确的数据类型。

4）准确性：观察变量的分布图，结合挖掘机技术知识，观察数据是否符合实际。

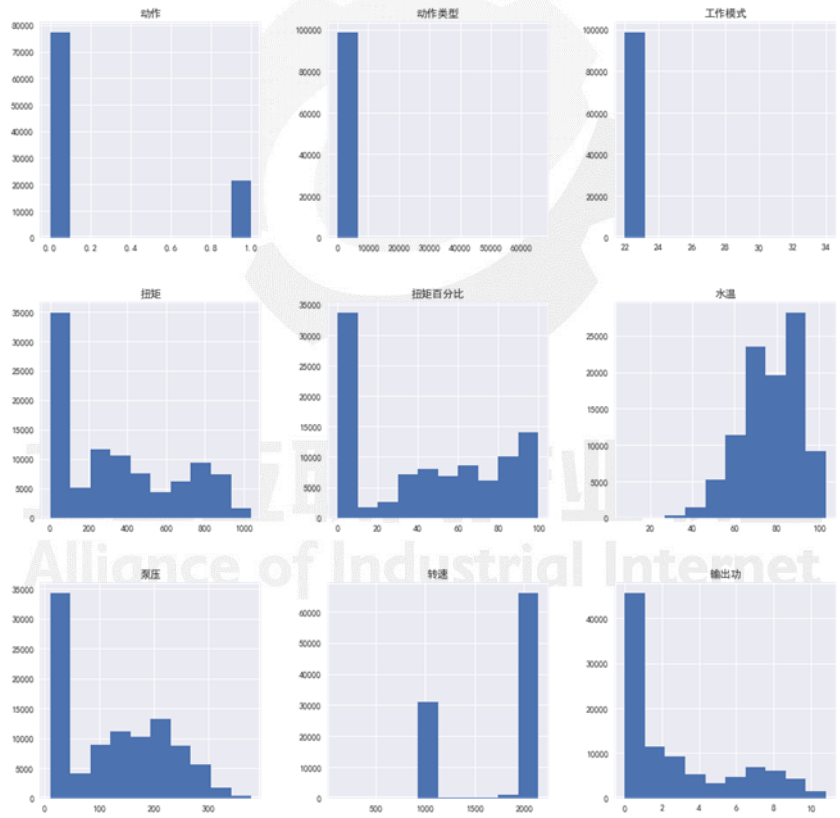


图 4 参数分布直方图

5）唯一性：数据每分钟传入的数据有且只有一个。

6）关联性：结合挖掘机工程技术知识，根据变量之间的相

互关联、约束等条件，检查数据的规范性。

3、数据准备

建模数据需选取挖掘机正常状态下的实时数据。根据正常状态下的数据建立一个预测模型，由这个模型去预测状态未知情况下的泵压数据。

（1）数据预处理

该阶段主要将采集到的原始数据经过校验、处理转变成可以用于建模的干净、完整的数据。首先需要校验每个字段的数据类型是否符合逻辑。再看其取值范围是否符合实际，再经过一系列处理将原始数据转换成我们所需要的数据。

表 2 数据的描述性统计

	mean	std	Min	25%	50%	75%	max
转速	1697.89	476.75	113.50	1002.6	1997.4	2042.0	2144.80
扭矩百分比	44.15	35.61	0.00	4.50	44.50	78.50	100.00
水温	77.56	13.59	8.08	68.58	77.18	89.84	103.21
扭矩	345.76	308.72	0.00	38.97	289.54	620.58	1038.84
输出功	2.73	3.12	0.00	0.02	1.35	5.06	10.82
动作	0.22	0.41	0.00	0.00	0.00	0.00	1.00
动作类型	96.31	308.55	0.00	0.00	70.00	89.00	65535.0
工作模式	22.00	0.07	22.00	22.00	22.00	22.00	34.00
泵压	131.87	89.58	10.00	34.00	127.00	208.00	378.00

在进行异常值、缺失值处理前，先将数据根据机号按采集时间升序排序。

（2）异常处理

异常值判定方法：超出每个字段的实际取值范围的数据均看作异常值。

异常值处理方法：删除。

（3）缺失处理

1）删除空值占比>70%的列;

2）删除除空值外，只有一个值的列;

3）删除空值占比>70%的行;

4）连续型变量、离散型变量均使用向上填充方法填充缺失值。

（4）归约处理

连续型变量：采用 Zscore 标准化方法;

离散型变量：独热编码。

4、数据建模

（1）特征工程

1）数据初步筛选

选择跟液压系统相关的参数，且在工作状态中（即满足发动机转速>0 且扭矩百分比>0 等限定条件）的数据，这样一来所选数据均是有效的，能够反映挖掘机工作状态。经过处理后，仅剩 98777 行数据。

2）特征变换

在很多机器学习任务中，数据集中的特征取值并不都是连续数值，而有可能会是类别值。由于部分模型只支持数值型的数据作为输入，因此，我们需要提前将数据集中的类别型变量通过独热编码进行预处理。

独热编码即 One-Hot 编码，又被称作“一位有效编码”，它

将一个有 m 个可能取值的特征变成 m 个二元特征，并且这 m 个二元特征每次只有一个被激活。

3）特征组合

独热编码后，再结合挖掘机相关模式（P，E 等）信息将变量重组。

对于连续型变量，其内在也存在着一些关联关系，可以通过一些运算得到新的重要的参数，如扭矩、输出功等。

4）特征筛选

首先根据专业知识可知，与泵压关联度较高的参数有液压油温、液压泵流量、液压泵内泄量、发动机转速、扭矩、扭矩百分比等（其中液压油温、液压泵流量、液压泵内泄量等数据目前无法获得），剔除掉跟研究变量无关的干扰参数。

再根据特征的相关系数，对相关系数超过 85%的变量只保留一个。

图 5 特征相关系数热力图

最终我们选用的参数有：发动机转速、扭矩百分比、扭矩、输出功、模式 *P*、模式 *E*、动作、动作类型等共八个特征。

（2）算法介绍：XGBOOST

XGBOOST 是在 GBDT 基础上发展起来的，全名叫 Extreme Gradient Boosting，与 GBDT 相比有一定的改进。传统的 GBDT 算法在优化时只用到了损失函数的一阶导数信息，XGBOOST 则对随时函数进行了二阶泰勒展开，同时使用了一阶和二阶导数的信息。此外，XGBOOST 借助 OpenMP，能自动利用单机 CPU 的多核并行计算，大大提高了运行速度。其次，与 GBDT 算法不同，XGBOOST 支持稀疏矩阵的输入，并且，XGBOOST 集成学习框架自定义了一个数据矩阵类 DMatrix，会在训练开始时对训练集进行一遍预处理，从而提高之后训练过程每次迭代的效率，减少训练时间。

本案例采用 XGBOOST，训练数据为 2020 年 8 月 7 日~2020 年 8 月 31 日没有任何故障时的数据，80%作为训练集，20%作为测试集，采用交叉验证方法进行参数优化。

经过优化后，模型 R^2=0.79。同时得到了特征重要度：

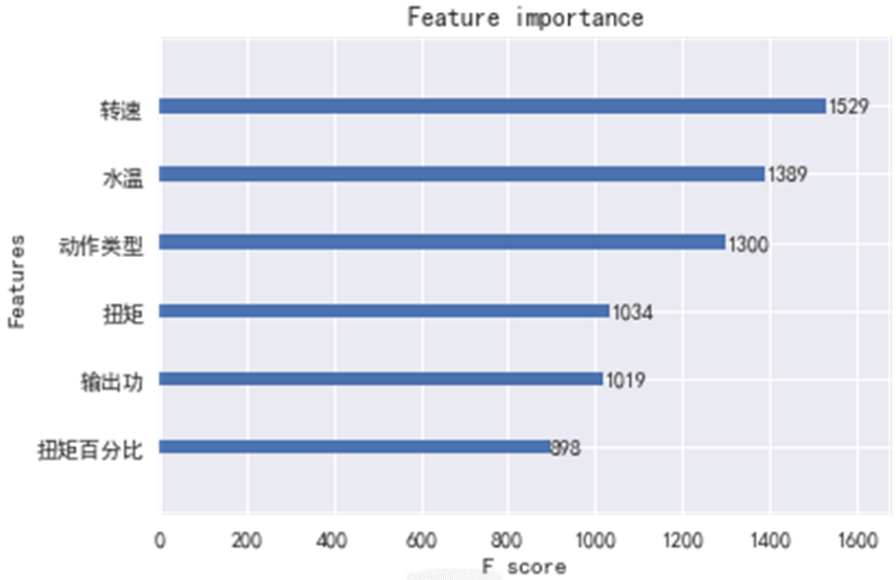


图 6 XGBOOST 特征重要度分布

5、模型验证

（1）验证逻辑

1） 根据上述模型获得泵压预测值;

2） 根据实际值与预测值计算每条数据的偏离度：$\triangle P/P$;

3） 计算不同偏离度（±5%，+10%等）每日占比;

4） 结合故障时间与偏离度占比趋势图分析;

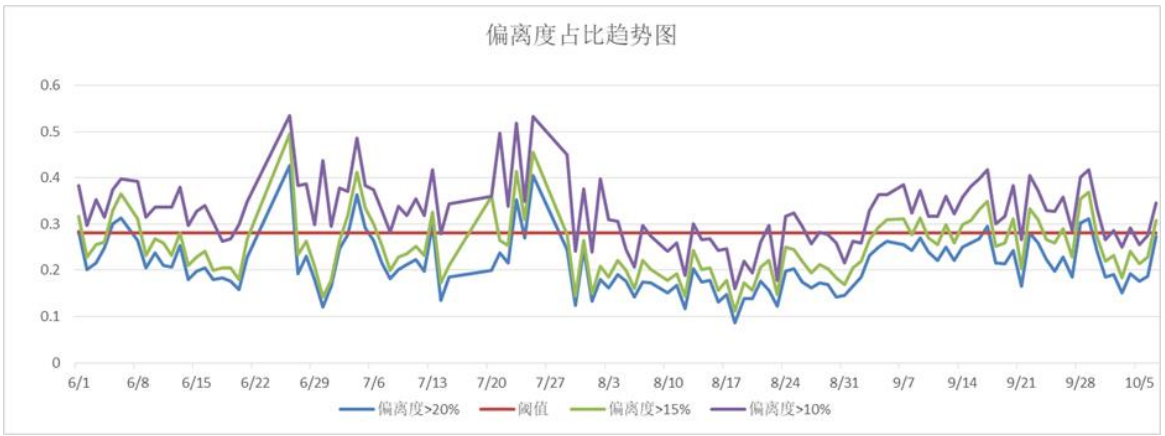


图 7 泵压偏离度趋势图

5）由上图可以得到一组模型设定值：w=20%，$threshold$=20%;

6）根据模型设定值，验证其他机号的预测精度;

7）调整模型参数，使其具有一定的泛化能力。

（2）方法评估

该项目最终的分析目标是预测液压泵是否故障，这是一个二分类问题。混淆矩阵是评价分类问题精度的一种方法，对于二分类问题，根据真实类别与预测类别的组合划分为真正例、假正例、真反例、假反例四种情况，令 *TP*、*FP*、*TN*、*FN* 分别表示对应的样例数。

在该案例中，我们假设故障报修时间前 5 天已经发生故障，这样我们得到的混淆矩阵如下表所示:

表 3 预测结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	12	6
反例	23	74

查准率 $P=TP/(TP+FP)=12/(12+23)=0.34$

查全率 $R=TP/(TP+FN)=12/(12+6)=0.67$

结论：结果显示，样例中的 3 次液压泵系统故障均被预测到，但是误报情况也时有发生，因此今后需要在判定方法上进行优化，减少误报。

（三）实施效果

1、模型部署

（1）模型部署的自动化

考虑到实际中机器的型号不同，其参数、性能也有一定的差别，在模型存储方面，我们将不同型号的车辆单独存储，在运行中，会自动调用对应的模型去运行结果。在自动化部署方面，我们实现了批处理程序的自动化，减少了人的干预，节省了人力时间成本。

（2）实施和运行中的问题

在实施和运行中普遍面临一个问题就是：建立分析模型所用的数据和运行中所用的数据存在差异。导致差异的原因包括：数据质量问题、精度劣化问题、范围变化问题等。针对这些问题我们将会在数据提取阶段检查数据质量，后续会根据预测结果对模型做持续的优化。

（3）问题的解决方法

针对数据质量问题，根据实际情况采取限制应用范围的方法，即当某机器某天的采集数量过少时，停止计算当天的模型并备注。

针对精度劣化问题，我们采用不定期地重新修正模型的方法，实现模型的自动修正。

（4）部署后的持续优化

要想技术有生命力，模型运行过程中就要进行持续优化。优化的内容包括精度的提高、适用范围的扩大、知识的增加等。

模型精度很大程度上取决于数据的质量。特定数据的质量往往取决于基础的维护和管理水平，而在某些特定项目中的使用到的数据其质量往往很差。因此，对于模型所用到的原始数据、故

障数据等的规范化、标准化是我们优化过程之中的重中之重。

2、应用效果

模型部署后，我们重点监控了 6 台模型预测为故障的挖掘机，并进行了现场派工检查。经过调查发现：其中 3 台挖掘机液压泵无任何异常，且用户也没有其他问题反馈，另外 3 台挖掘机液压泵虽无异常，但调查发现一些不影响挖掘机正常使用但跟液压泵有关联的一些异常表现，如憋车、动作慢等。总体来说，目前的实际应用效果还有待进一步提高，具体原因可能有以下几点：

一是目前数据缺少与泵压相关的一些重要参数，尤其是液压油的一些参数，如油黏度指数；

二是实际问题往往不是一个单一的算法可以解决的，需要多个相关算法合理的搭配组合，再结合机理模型进行综合考虑；

三是液压泵故障是一个复杂的问题，液压泵故障会导致泵压降低，但是反过来泵压降低也有可能是其他零部件故障或操作异常等导致的。

由于上述客观问题的存在，我们只能在现有数据条件的基础上进行有限的优化，比如：扩大样本量、试验不同的模型组合、优化异常判定模型等，以此来减少预测结果误报率，提升预测结果准确率。

二、案例 2：数控加工工艺参数优化—华中数控

（一）案例背景

数控机床被称为“工业母机”，是传统机床与数字控制技术相结合形成的机械电子一体化产品。数控机床具有稳定、高效、灵活等各种优异性能，开创了传统机械向机械电子一体化发展的先河。数控机床等数控装备是生产高新科技装备和尖端产品的必要工具，可以有效地提高生产效率、减少工人数量，实现自动化、智能化生产，在很大程度上减少了人员和成本投入。在当今以产品更新迭代快速、大批量生产、人员成本逐渐升高为特点的工业时代，各类数控装备是实现先进制造技术的关键。因此数控机床成为了国民经济和国防建设发展的重要制造装备。

在数控加工领域，对加工质量、加工效率与加工成本的控制能力是衡量加工能力强弱的指标，而如何提升这种能力，亦即针对于工艺参数优化的研究，在数控加工领域占据着重要地位。在工艺参数优化问题中，往往需要同时关注多个优化目标，效率、质量与加工成本的优化需要同时被解决。而多个优化目标之间可能彼此矛盾，亦即在提升加工效率的同时可能会造成加工质量的降低，提升加工质量的同时又会造成加工效率的降低与加工成本的提高，如何有效地实现多个目标的同时优化，是当前工艺参数优化领域所面临的一项重要任务。

对加工过程进行优化，不可避免地要对加工过程进行建模。加工过程模型一般包含有几何模型及物理模型两种，由于物理模型往往能更直接地反映加工过程的力、热特性，因此加工过程物

理模型被广泛应用在工艺参数优化中。在大多数的研究中，常常通过设计实验的手段来构建加工过程物理模型，通过设计实验可以在较短的时间内获得较为全面的特征数据以加快建模的进程。但设计实验也融入了人为的主观因素与局限性，因此必然无法涵盖所有的加工工况。与设计实验不同的是，日常加工任务的加工过程数据实际上涵盖了大量的工艺信息，这部分信息中拥有设计实验所无法涵盖的特征数据，是非常具有价值的工艺信息。加工过程数据具有数据量大、与工艺系统工艺特性强相关以及加工场景日常等特点。

本案例以特定的工艺系统为研究对象，以工艺参数提取、基于指令域分析的数据融合策略、优化后的神经网络模型以及多目标优化算法，实现了在特定工艺系统中的加工过程建模及铣削粗加工进给速度优化。在满足生产需求的前提下，选择合理的加工参数，充分发挥机床的性能，对于生产效率提升、改善刀具寿命和降低工艺人员工作强度和产品的成本，提升企业竞争力有重要意义。

（二）解决方案

1、业务理解

在进行数控加工编程时，加工工艺参数（如切深，切宽，主轴转速，进给速度等）的确定至关重要，它们影响着零件的加工质量、效率、机床和刀具等制造资源的寿命等。当工艺参数选择过于保守，机床和刀具等制造资源的效能会产生严重浪费，影响加工效率和经济性；相反，过于激进的工艺参数会使加工中产生

诸如机床颤振、刀具磨损、加工质量低下等异常情况，严重时甚至会带来制造资源的损坏，产生重大生产事故和经济损失。

由于在传统的铣削加工过程中，不同的指令行所对应的材料去除率是不同的，如下图所示，在优化前，指令行 1、2、3 对应于切削厚度为 h1、h2、h3，其中 h3<h1<h2，但是往往采用了相同的进给速度 F，这导致主轴功率波动较大，且在切削厚度较小的指令行的切削效率未得到充分的挖掘，为了降低主轴功率波动，提升切削的效率，可以通过将材料去除率较小的指令行的速度提升，如指令行 1 和指令行 3，而将材料去除率较大的指令行的速度降低，如指令行 2，最终在得到稳定的主轴功率的同时，提升加工效率。

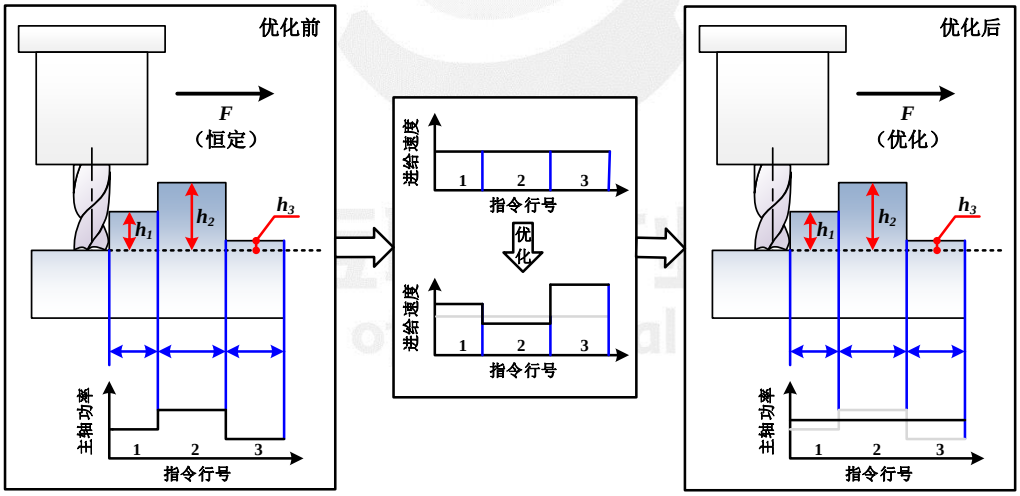


图 8 进给速度优化原理

在上述原理的基础上，本案例提供两种解决方案。

（1）方案一：基于指令域“心电图”的加工工艺参数优化

此种优化方案，需要对零件进行试切加工，并采集加工过程中的切削负载数据，基于采集的实际负载数据对加工工艺参数进

行优化。需要多次实际加工进行迭代优化，以得到最终优化后的工艺参数。

（2）方案二：基于加工过程主轴功率模型的加工工艺参数优化

此种优化方案，在优化前需要建立切削加工工艺参数与切削负载之间的映射关系（即模型）。基于建立的模型，当有新的零件需要加工时，无需试切加工，即可利用模型对加工过程的切削负载进行预测，基于预测的切削负载数据对加工工艺参数进行优化。此方案是基于模型预测结果离线迭代的切削工艺参数优化。

两种方案有着相同的优化目标：在优化加工效率的同时、降低切削功率波动。

两种方案各有优点：方案 1 不需要事先建立模型，方法比较简单，但是优化前需进行试切加工，主要适用于大批量工件的工艺参数优化；方案 2 需要通过试验建立模型，方案比较复杂，但是优化时却不需要试切加工，主要适用于单件或者小批量产品的加工优化。

2、数据理解

G 代码是数控加工过程的关键，它涵盖了加工过程的主要信息并指导实际加工。传统的数据分析方法大体可分为时域分析与频域分析两种手段，而在数控加工领域，仅仅依靠上述两种手段无法实现对数据的有效利用与分析。基于此，本案例应用基于指令域的大数据分析方法，该方法以加工过程数据与 G 代码指令数据之间的对应关系为基础，将指令行作为数据分析及统计的基

本单位，从按时间分析、按频域分析转变为按指令行分析，进而使得对数控加工过程数据的分析更加科学、准确与全面，对分析结果的应用更加有效。

所谓指令域大数据，是指将数控加工过程通过制造资源、工作任务和操作状态进行关联描述。

1）制造资源。数控机床的制造资源（简称“MR”）是指执行加工任务所需的物理系统部件(包括主轴、丝杠、导轨、轴承、电机、刀具等工艺系统数据）以及加工过程中的环境因素（如温度、振动等），主要是数控机床的属性数据、参数数据等。

2）加工任务。数控机床的加工任务（简称“WT”）是指数控机床在给定的制造资源条件下完成特定的工作任务，即数控机床的任务数据。

3）运行状态。数控机床的运行状态（简称“*Y*”）是指在工作过程中产生的工作质量和效率等可以表达的特征参数，包括主轴功率、扭矩、振动、进给轴轮廓误差等电控数据，即数控机床的逻辑数据和状态数据。

通过式（1）建立制造资源数据、工作任务数据与运行状态数据的映射关系 $Y = f(WT, MR)$，输入为制造资源 MR 和工作任务 WT，输出为对应的操作状态数据 Y。需要指出的是，式（1）的表现形式为一组大数据。

$$Y = f(WT, MR) \tag{1}$$

数控机床的加工任务通过加工零件的 G 代码程序进行定量描述。在已定的制造资源条件下，G 代码程序及其所包含的加工

指令共同描述工作任务数据，如被加工零件的形状特征、尺寸、加工模式等。G 代码通过指令行号定义指令执行顺序，构成加工系统的运动轨迹和加工模式。综上所述，G 代码可以解释为以插补周期为间隔、由指令行号定义执行顺序的时序数据，显式或隐式地描述指定时刻（或指令行号）的加工参数，如刀具、夹具、主轴转速和进给速度等，以及当前时刻的振动、温度、电流、功率等运行状态数据。因此，G 代码指令行号可以作为连接制造资源、工作任务与状态数据的“索引”，构建数控机床的指令域大数据，数据结构可通过下图表达。

$$
\begin{array}{ccccccc}
M & \rightarrow & m_0 & m_1 & \cdots & m_N \\
 & & \downarrow & \downarrow & \downarrow & \downarrow \\
D^1 & \rightarrow & d_0^1 & d_1^1 & \cdots & d_N^1 \\
D^2 & \rightarrow & d_0^2 & d_1^2 & \cdots & d_N^2 \\
\cdots & \rightarrow & \cdots & \cdots & \cdots & \cdots
\end{array}
$$

图 9 指令域大数据结构

其中，$m_t = (MR_t, WT_t) \quad t = (0, \cdots, N)$；

MR 为制造资源，包括刀具、夹具等；

WT 为加工任务，主要以 G 代码形式表现；

$d_t \quad t = (0, \cdots N)$，为状态信号，包括位置环、速度环、电流环及外部感知数据。状态信号数据根据 G 代码指令行号被有序“分组”，从而实现制造资源、加工任务与运行状态数据的映射，成为“加工换轨”这一重要加工事件的标识。指令域大数据从表面来看只是强调自身是基于 G 代码行号的关联数据，但实际上隐

含有时间属性。

前面已经提到本案例所提出的两种解决方案均是建立指令域大数据的基础上实现进给速度优化目标。

方案一基于指令域“心电图”的工艺参数优化，需在首次加工过程中实时的采集加工过程中的程序行号、负载数据（主轴功率或主轴电流），X 轴、Y 轴、Z 轴指令位置和指令速度，以此建立指令域“心电图”以及进给速度优化模型。

方案二选用可从数控系统内部直接获得的且能够表征加工过程物理状态的主轴功率代替铣削力，并以日常加工任务数据生成样本数据，进而应用神经网络建立主轴功率预测模型。该方案以进给速度、主轴转速、切宽及切深建立针对特定工艺系统（固定的机床、刀具及毛坯材料）的加工过程主轴功率模型并将特定工艺系统的工艺特性拟合在模型当中。

3、数据准备

方案一中使用配备有华中 8 型数控系统的台群机床 T-500，如图 10-a 所示，加工零件为手机外壳，并采用多把刀具进行加工。

方案二中使用华中数控公司测试车间的 Z540-B 钻攻中心（配备有华中 8 型数控系统，如图 10-b 所示）作为实验机床，使用直径 8mm 的三刃立铣刀作为加工刀具，毛坯材料为 7075 航空硬铝。

两种方案中使用华中数控数据采集软件 SSTT 采集，实验测量了不同工艺参数组合下的主轴功率并对行号以及各轴的实际

速度，通过 SSTT 数据采集软件与数控系统通讯，利用数控系统的开放接进行了采集，采样频率为 1kHZ。如图 11 所示为 SSTT 数据采集软件的采样界面。

（a）T-500　　（b）Z-540B

图 10 实验机床

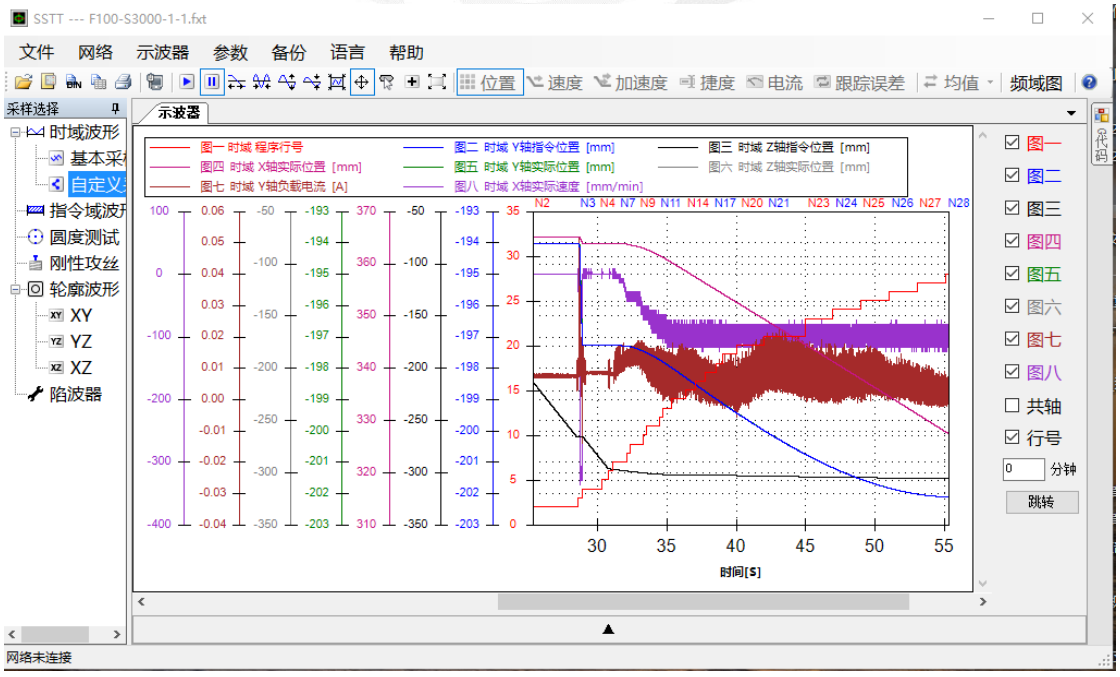


图 11 SSTT 数据采集软件采样界面

在实验过程中，机床加工过程的运行数据根据下图所示的方式进行采集，数控装置的总线及寄存器，会在每个控制周期更新

数控装置下发给伺服控制器的指令数据，以及光栅尺、编码器等反馈的实际数据。将 PC 电脑与数控装置通过网线连接，通过与数控系统适配的数据采集软件 SSTT，即可采集得到数控机床的实时运行数据。

图 12 数据采集演示

通过数据采集软件 SSTT，设置所需采集的数据类型，即可实时采集机床进给系统的运行数据，数据采集周期为 1ms。其与数控系统通信并实时地采集包括指令行号、各轴指令位置、各轴实际位置、各轴指令速度、各轴实际速度、各轴加速度以及各轴功率或电流在内的各类加工数据。

对于指令位置数据，在采集时需要记录工件坐标系在机床坐标系下的坐标值，以便将采集得到的机床坐标系下的指令位置数据转化为在工件坐标系下的指令位置坐标值。

（1）数据的预处理

机床工艺系统的运行数据记录了加工过程中的特征信息，反

映了系统的输入与输出之间的对应映射关系。但是，机床工艺系统的运行数据种类繁多，成分复杂，包含许多无用的数据及干扰，且数据之间的量纲往往不同，我们所关心的特征可能会由于幅值较小，在原始数据中仅占很小的部分，直接对原始数据进行分析，不容易找到数据之间的关联关系。因此，在建立模型之前，需要对运行数据进行分析处理，找出数据之间的因果关系，以及数据的不同特征对实际响应的影响效果，尽量分离出原始数据中对建模无用的数据，有效避免各种干扰，放大特征，便于后续的建模过程。

本案例中，方案一采用的数据预处理方法主要如下:

①工艺参数提取。从 G 代码信息中，直接提取出主轴转速、进给速度等工艺信息。

②数据滤波。通过SSTT 软件采集到指令数据中，主轴功率、实际速度等含有高频噪声成分，或有其它噪声干扰，需去除噪声，提升数据有效性。

③指令域分析。将每一行的工艺参数与实际的响应数据在指令域上进行一一映射，可用于分析实际加工状态。

本案例中方案二采用的数据预处理方法及流程主要如下:

①工艺参数提取。以毛坯几何信息或 STL 模型作为毛坯 Z-map 离散建模的输入，对 G 代码中的工艺信息进行读取并进行刀位点的离散。基于离散的刀位点生成刀具扫掠体对材料去除量切宽、切深进行提取并将工艺信息在指令域表示。

②数据的无量纲化。对数据进行归一化操作，转化成为“纯

量”数据，各指标数据处于同一数量级，便于不同单位或量级的指标进行比较和加权，加快模型训练过程中梯度下降速度，改善模型的收敛性。

③数据的滤波。采集得到的数据，如实际速度、实际位置、跟随误差等均存在着大量高频的波动值，其波动幅度小，频率高，规律性小。因此需要对数据进行滤波，去除掉高频波动干扰，再用于模型的训练。

④数据对齐与融合。工艺参数的提取按照距离进行分隔，SSTT 的数据采集频率为 1kHZ，需要将二者在刀位点上细致对齐，才能准确反应每一刀位点对应工艺参数与响应数据，保证生成样本的有效性。

（2）输入输出的确定

本案例中方案一通过主轴电流的变化可以反映进给速度的变化，而且主轴电流的变化还能够反应机床的负载状态，因此选用主轴转速、进给速度、主轴电流材料去除率作为模型输入，以进给速度作为模型输出。

本案例中方案二以主轴功率来表征切削力并进一步地反映实际加工状态。实际影响切削力的因素包括有切削液、毛坯材料硬度、刀具材料类型、刀具直径、顺逆铣、切宽、切深、刀具磨损情况、刀具前倾角、后倾角、刀具刃数以及机床及主轴的特性等。考虑到工艺响应与工艺系统工艺特性相关，因此本案例基于特定的工艺系统建立主轴功率预测模型，将特定工艺系统（刀具、毛坯材料及机床）的工艺特性拟合至模型当中。选取进给速度、

主轴转速、切宽以及切深作为模型输入，以主轴功率值（经均值滤波）作为模型输出。

4、数据建模

与方案一对应的是基于指令域“心电图”的加工工艺参数优化，与方案二对应的是基于加工过程主轴功率模型的加工工艺参数优化。

（1）基于指令域“心电图”的加工工艺参数优化

在进行工艺参数优化时，单次切削时间、零件表面质量和刀具寿命是首先要考虑的几个方面。减少单次切削时间无疑会带来经济效益，但是在减少单次切削时间的同时会导致零件表面质量和刀具寿命在一定程度上都受到损害。因此，此方案中工艺优化主要有三个目标分量分别是单次切削时间、零件表面质量和刀具寿命。

方案一主要的研究对象是粗加工，在粗加工中主要关注的是加工的效率问题，而对零件的表面质量并没有太大要求。所以在优化时，目标函数首先是单次切削时间，要通过优化加工参数减少单次切削时间。其次是刀具寿命，通过对工艺参数的寻优使得加工过程中刀具的负载趋于均衡，有利于保护刀具寿命。

在以提升效率为目标的粗铣加工工艺参数优化过程中，刀具寿命是最可能受到不利影响的因素。在本方案中，以主轴负载电流的变化来反映进给速度、材料去除率的变化。优化的限制条件也应该反映到主轴负载电流上。具体来说，将原始 G 代码加工时产生最大电流时的切削参数作为最大限制条件，即机床性能得到

充分发挥是能产生的电流最大值，优化后的最大电流应小于原始G代码加工产生的最大电流。在这约束条件下，材料去除率是在机床和刀具所能承受的最大范围内。原始 G 代码加工时产生的最小电流也应该作为优化的限制条件之一。优化后 G 代码加工产生的电流应大于原始 G 代码产生的最小电流，因为若优化后主轴电流变小则意味着进给速度或是材料去除率变小，这样就不能达到提高加工效率的目的。另外在优化进给速度的时候要注意到进给速度的变化过大会造成机床的振动、表面质量不佳等情况。

优化约束条件:

1）优化后的主轴电流最大值应小于原始G代码的主轴电流最大值;

2）优化后的主轴电流最小值应大于原始G代码的主轴电流最小值;

3） 进给速度平滑过渡;

4） 优化后的 G 代码在加工后的零件满足表面质量要求。

（2）基于加工过程主轴功率模型的加工工艺参数优化

人工神经网络是模仿动物神经中枢而建立的函数拟合模型，其具备良好的非线性拟合能力。人工神经网络是由多个人工神经元联结而成，单个神经元一般以 $x_1 \sim x_n$ 为输入向量，$w_1 \sim w_n$ 为输入向量各分量所对应的权值，b 为偏置，f 为激活函数，θ 为阈值，t 为输出，则 $t = f\left(\sum_{i=1}^{n}(w_i \times x_i + b) - \theta\right)$。经典的神经网络一般由多层神经元所构成，每一层神经元都具备输入及输出。常见的多层前馈

神经网络一般包含输入层、输出层以及隐藏层，此种网络亦被称为多层感知机。神经网络的学习是通过训练集数据对各层权重及神经元的阈值进行反复校正，直至网络输出的目标值与实际目标值的相符程度达到一定的精度范围，进而完成训练。

前文对机床的加工过程数据进行预处理，对指令数据与工艺信息数据之间进行对齐融合、特征的提取与分析，确定了预测模型的输入、输出。

反向传播算法是一种常见的神经网络训练算法，它以网络的计算误差为依据逐层更新权重向量，反复执行正向传播与反向传播两个过程直至计算误差达到允许范围内而完成学习。本案例即采用基于反向传播算法的 BP 神经网络来建立三轴铣削粗加工主轴功率预测模型。本模型的结构如下图所示。

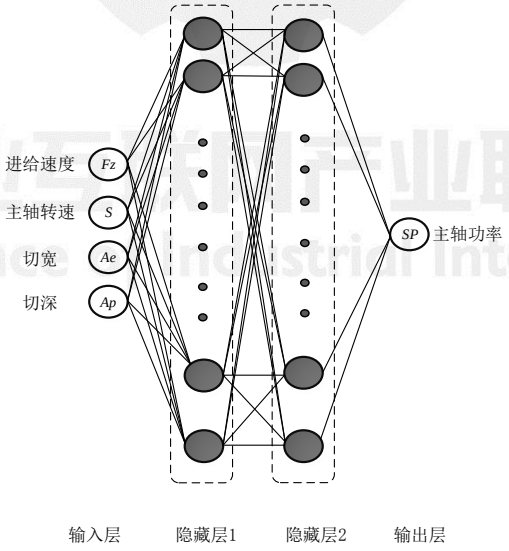


图 13 神经网络结构

在训练模型的过程中进行实际测试，如图 13 所示，根据训练预测效果最终选用网络结构为输入层—隐藏层—隐藏层—输

出层，对应节点数为 4-10-10-1 的神经网络模型作为最终的训练模型。

5、模型验证

（1）基于指令域“心电图”的加工工艺参数优化

案例一中所加工的零件为东莞劲胜精密组件股份有限公司代加工的某品牌手机模的二夹，工艺阶段主要为粗加工。加工的零件如下图所示。加工所用的 G 代码包括 1 个主程序和 15 个子程序，一共使用了 12 把刀，刀具半径从 1mm 到 10mm 不等。主轴转速的变化范围为 10000r/min 到 18000r/min，进给速度范围为 300mm/min 到 6000mm/min，具体信息如下表所述。在确定加工零件、制造资源和 G 代码后进行实验并采集电流数据，采样周期为 1kHz。

图 14 手机模二夹零件图

表 4 实验条件

序号	程序名	刀号	刀具直径(mm)	转速(r/min)	进给速度(mm/min)
1	O5001	T1	D10	10000	5000-6000
2	O4002	T2	D6	10000	3000-4500
3	O4003	T3	D6	10000	2500-3000
4	O5004	T4	D4	12000	2500-3000
5	O4005	T5	D3	12000	500-1300
6	O5006	T6	D4	12000	2000-3000
7	O5007	T7	D2	14000	1500-2000
8	O4012	T12	D1.5	18000	1500
9	O4008	T8	D1.6	16000	1500-2000
10	O4009	T9	D1.6	16000	1500-2000
11	O4010	T10	D1	16000	300-800
12	O5011	T11	D4	15000	4000

以程序行号为横坐标，指令域的主轴电流均值为纵坐标绘制主轴电流。其中红色为原始 G 代码加工时的电流，蓝色为优化后 G 代码加工时的电流。从优化前后主轴电流对比图中可以直观的看出优化后电流波动小于优化前电流波动，加工的主轴电流趋于均衡。并且在优化后电流有整体的提升。优化前单次切削时间为 7 分 49 秒，优化后时间为 7 分 4 秒，效率提升 9.59%。

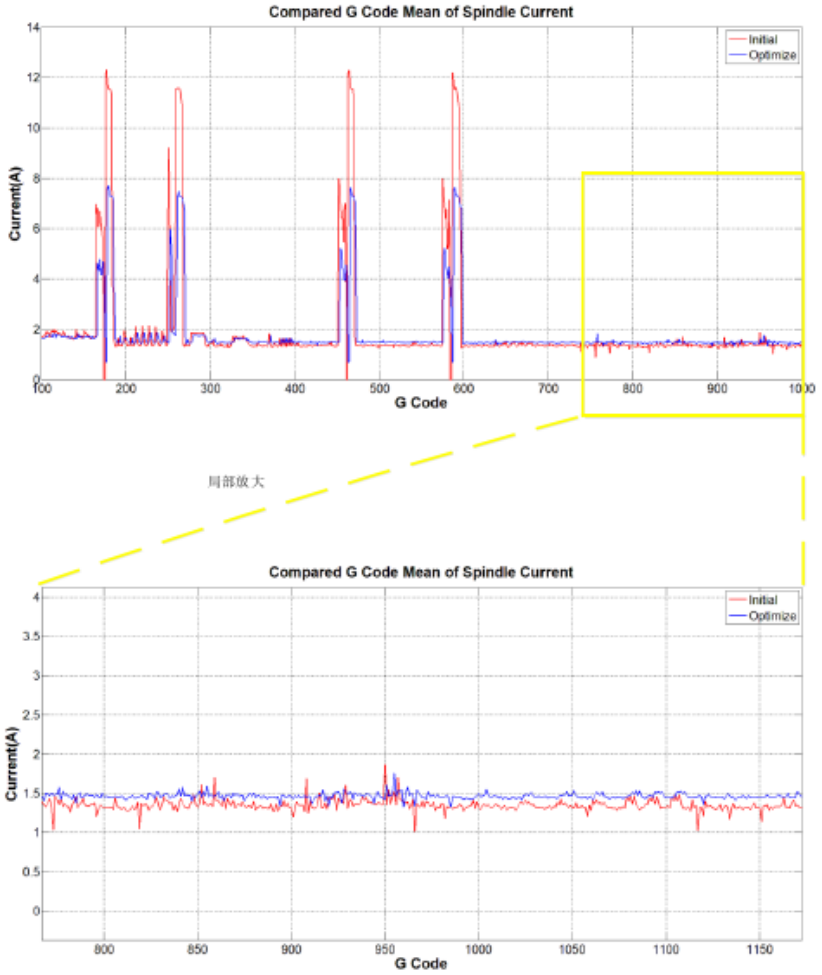
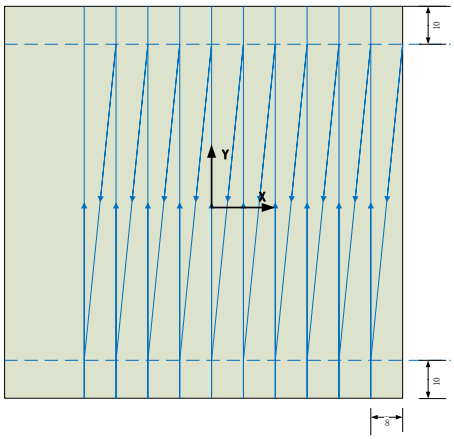


图 15 优化前后主轴电流对比

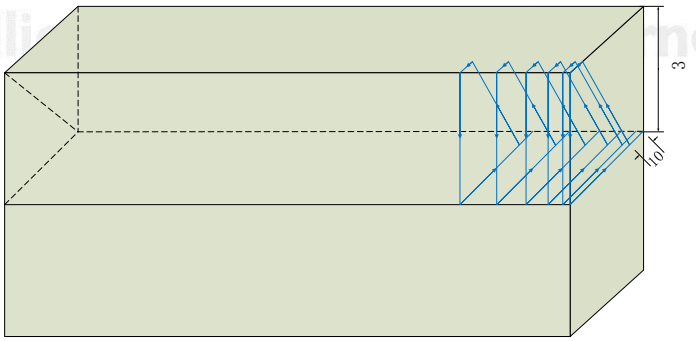
（2）基于加工过程主轴功率模型的加工工艺参数优化

为验证上述所构建的模型预测的精度，设计了变切宽铣削实验、变切深铣削实验，生成测试集，并在多次加工过程中采用不同进给速度下进行实验。两种类型的验证实验加工轨迹如下图所示。



加工次数	主轴转速（r/min）	进给速度(mm/min)	切深mm	切宽mm
1	4000	200、400、600 800、1000、1200、1400、1600、1800、2000	0.4、0.8、1.2、1.6、2.0	0~8连续变化
2	5000			
3	6000			
4	7000			
5	8000			
6	9000			
7	10000			
8	11000			
9	12000			
10	13000			

图 16 变切宽实验



加工次数	主轴转速r/min	进给速度mm/min	切宽mm	切深mm
1	4000	300、500、700、900、1100、1300、1500、1700、1900、2100	1、2、3、4、5	0~3连续变化
2	5000			
3	6000			
4	7000			

5	8000
6	9000
7	10000
8	11000
9	12000
10	13000

图 17 变切深实验

将测试集输入至主轴功率预测模型，其预测结果见下图，分别对误差及预测结果进行放大。从误差图中可以看到误差值主要在 5%附近波动，个别样本点的预测误差最低可达 2%，最高会达到 15%，但模型整体的预测误差为 4.91%。

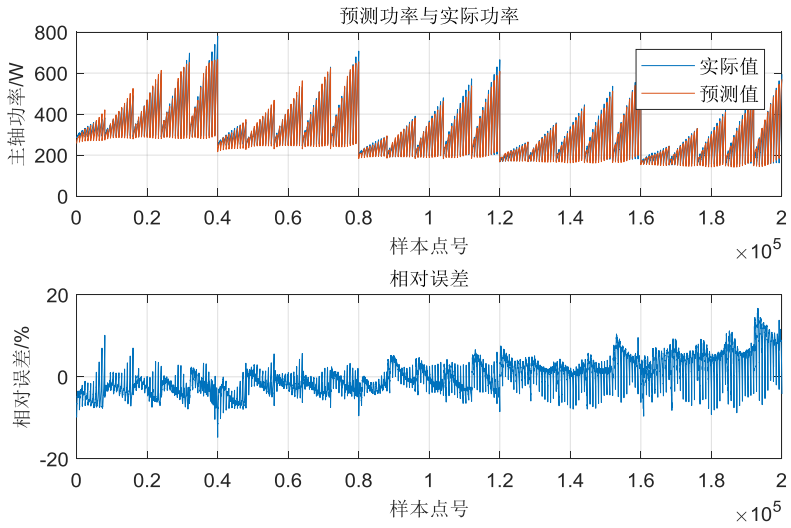


图 18 主轴功率预测模型预测结果

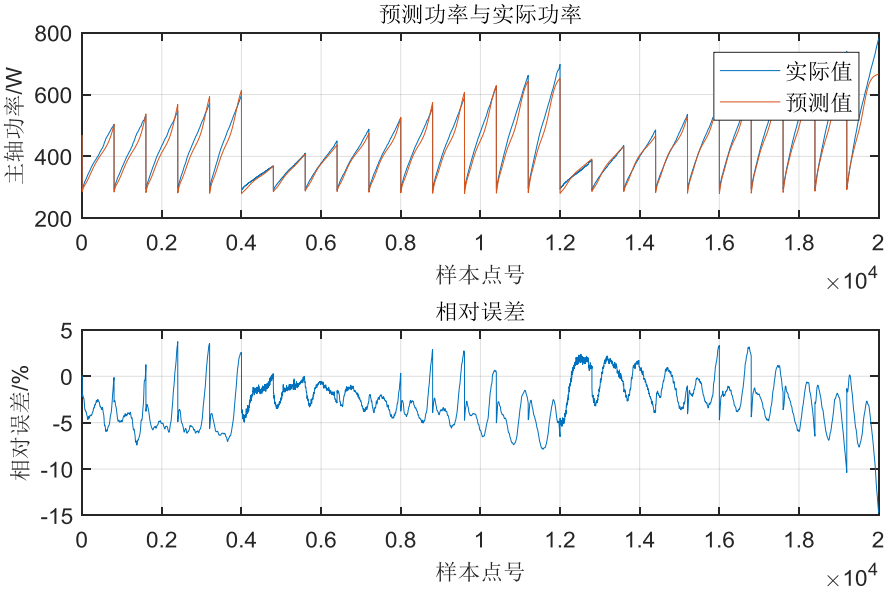


图 19 预测结果与误差放大（第 20000 至 40000 样本点）

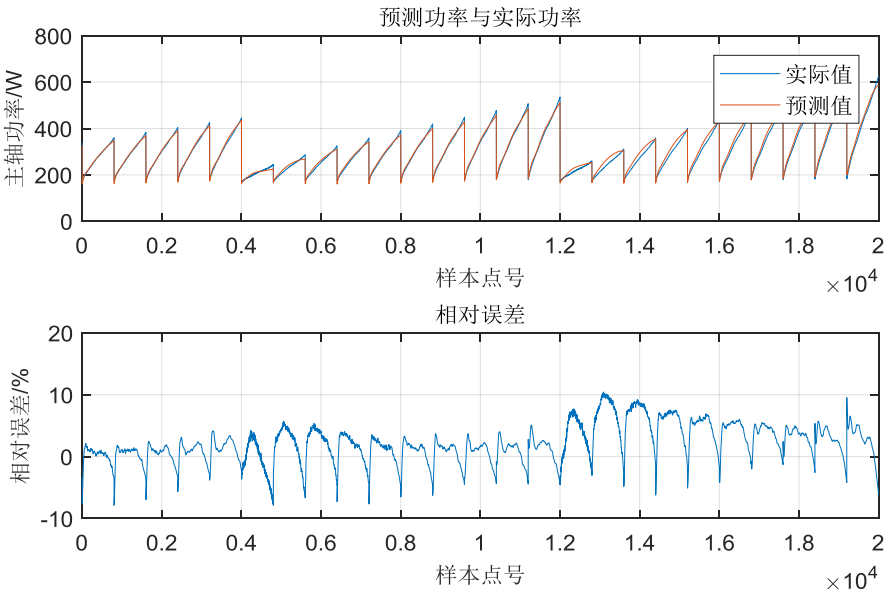


图 20 预测结果与误差放大（第 140000 至 160000 样本点）

（三）实施效果

1、基于指令域“心电图”的加工工艺参数优化

前面已经提到本案例所提出的方案一：基于指令域“心电图”的粗铣加工工艺优化是建立在试切（首件加工的程序行号，主轴

功率，指令位置，指令速度）基础上的进给速度优化方法，因此不同于传统的工艺优化方法在加工过程中实时的采集加工过程中的负载数据，并实时调整进给率。本案例提出的方法基于华中数控的双码联控技术，通过离线生成加工优化代码（第二加工代码，i 代码），并于 G 代码共同控制机床运动，实现加工过程的自适应优化，其过程如下图所示。

实际加工过程中，通过 i 代码修正原始 G 代码程序中指定行程序段的进给速度，即 F 值。能够有效平衡粗加工过程中切削负载，降低主轴电流的波动，在机床切削能力范围内合理地提高零件的加工效率。

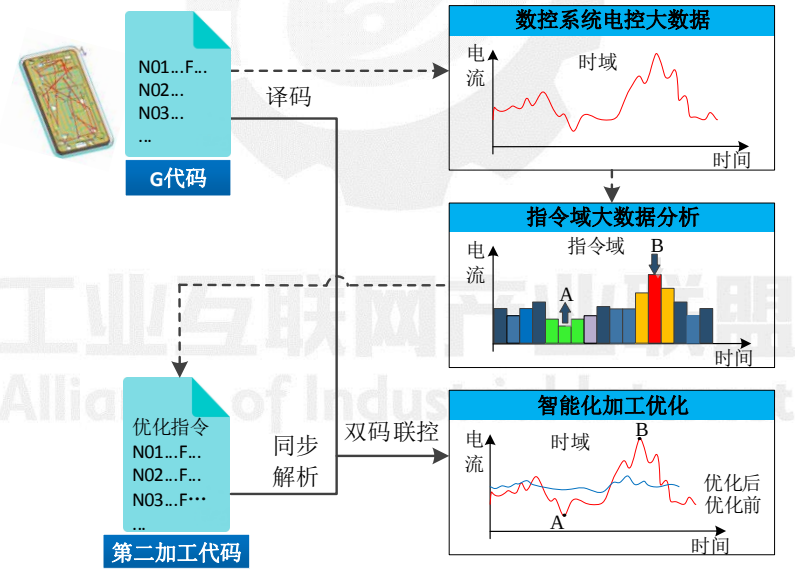


图 21 基于双码联控的工艺参数优化原理

下图中表示第 41 段原始 G 代码程序，按 i 代码中指定的 F1000.0 的进给速度运行。

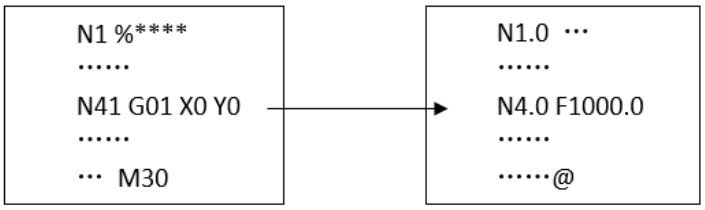


图 22 工艺参数优化的 i 代码编程示例

（1）工艺优化软件开发

基于上述原理，开发了在 Windows 环境下运行的工艺优化软件。工艺优化软件运行的环境如下图所示，使用网线将数控装置与 PC 机进行物理层的互联，通过 TCP/IP 协议构建 PC 机与数控装置的局域网，实现 PC 机对数控系统内部数据的采集，利用工艺优化软件分析采集的数据，实现了对进给率的优化，生成数控系统可识别的优化后的第二加工 NC 代码，并与 G 代码一同控制，实现对加工过程的自适应控制与优化。

图 23 工艺优化软件运行环境

工艺优化软件的软件界面如下图所示

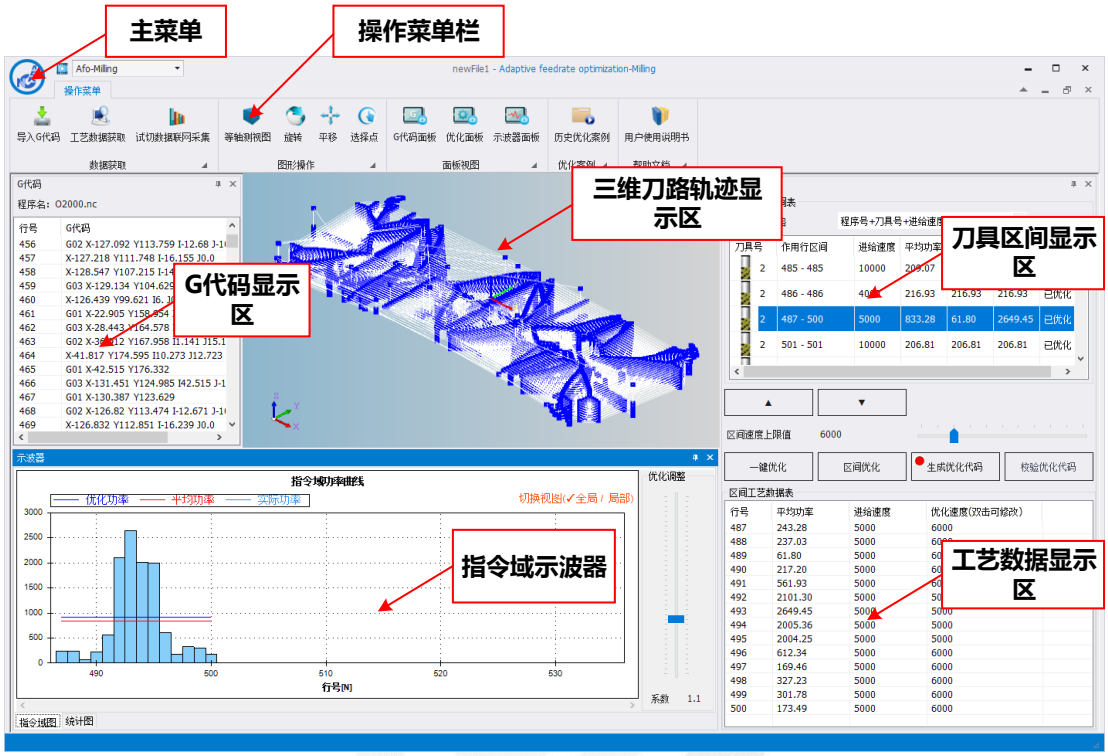


图 24 工艺优化软件的 HMI

（2）工艺优化软件的应用验证

为了验证本文提出的基于指令域“心电图”的加工工艺优化方法、双码联控技术及工艺优化软件在实际加工优化中的成效，选择了用户的实际加工零件进行了加工优化，待加工优化的零件为汽车轮毂，如下图所示。

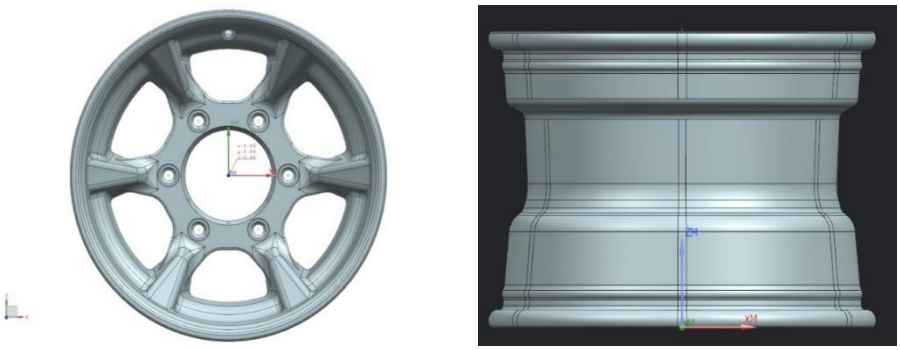


图 25 待加工优化的汽车轮毂件

加工用机床为 V1160L 立式加工中心，主轴最高转速 18000rpm，如图下所示，工件装夹如下图所示。

图 26 加工用机床 V1160L

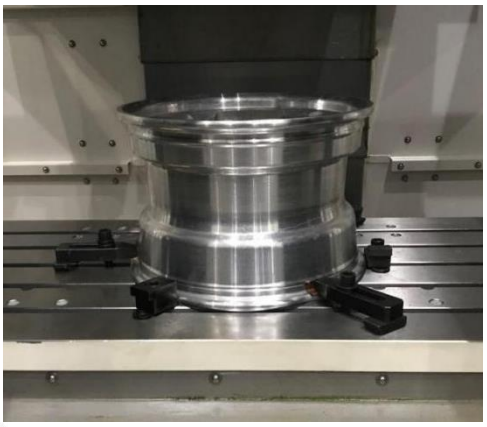


图 27 汽车轮毂装夹图

运用工艺优化软件，针对首件加工的数据对 G 代码进行优化，并生成优化后的 i 代码，软件界面如下图所示。

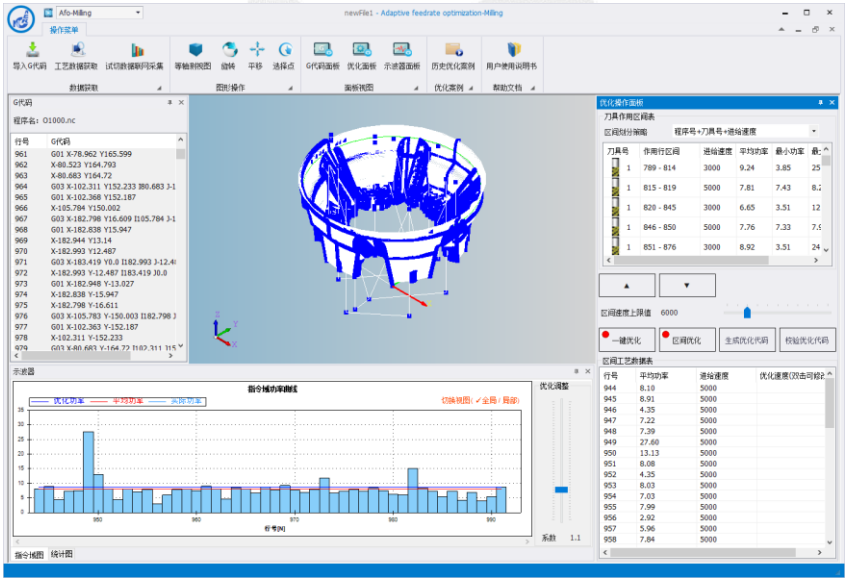


图 28 软件优化界面

生成的优化 i 代码与原始 G 代码对比如下图所示，从图中可以看到，在原始 G 代码中，进给速度值经常保持一个恒定值，对

于汽车轮毂加工这种切削余量变化很大，且毛坯不均匀的加工情况，采用固定的切削进给速度是不合理的，而优化后的 i 代码进给速度根据主轴负载的变化进行了调整。

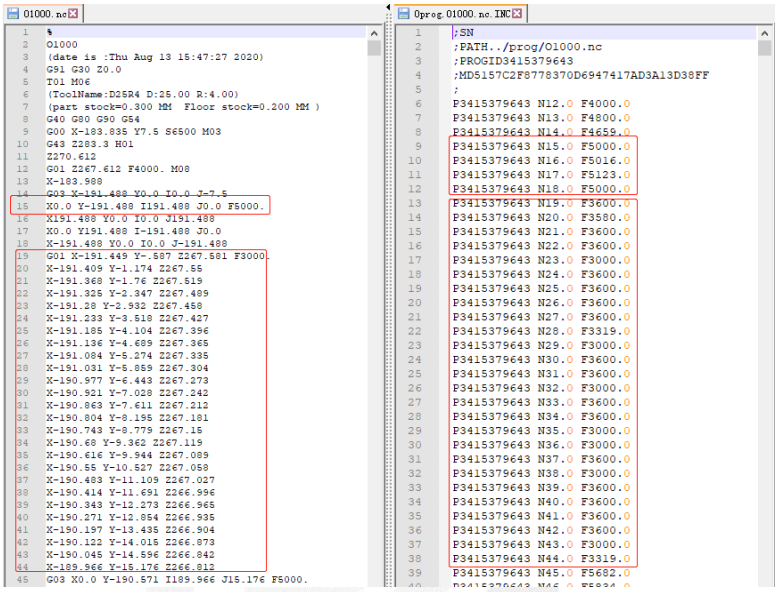


图 29 优化前后 G 代码与 i 代码对比

利用优化后的 i 代码导入华中 8 型数控系统，共同进行加工，粗加工完成后的零件如下图所示，从下图可以看出，优化完成后表面质量无瑕疵，满足客户要求。

图 30 汽车轮毂粗加工（优化后）

优化前后的加工时间如下表所示。

表 5 工艺参数优化前后对比

	加工时间	效率提升
优化前	1h23min	
优化后	1h17min	7.2%

从表中可以看出，采用工艺参数优化后，汽车轮毂粗加工时间由 1h23min 缩短为 1h17min，粗加工效率提升 7.2%，且加工质量满足要求，结果表明本课题提出的基于指令域“心电图”的粗铣加工工艺优化技术在实际加工中具有提升加工效率的作用，对于加工企业节省加工成本具有重要意义。

此外，该优化方法及软件在航空、航天、汽车、3C 电子等重点领域的重点用户进行了实际的加工应用，取得了良好的应用效果，提升加工效率 5~30%，以下是部分加工优化案例。

（a）运载火箭结构件

效率提升 11%

（b）航天结构件

效率提升：9.8%

（c）特种高性能发动机活塞

效率提升：5.2%

（d）汽车底盘零部件

效率提升：2¯5%

（e）3C 电子，效率提升：5¯30%

图 31 应用案例

2、基于加工过程主轴功率模型的加工工艺参数优化

基于本案例中方案二所开发的加工过程主轴功率模型，通过设计实验进行了验证。所开发模型具体部署环境如下：

实验用例：如下图所示；

实验环境：数控系统版本：HNC2.4；

机床：Z-540B 钻攻中心；

加工刀具及材料：三刃平底立铣刀，其直径为 8mm，毛坯材料为 AL7075；

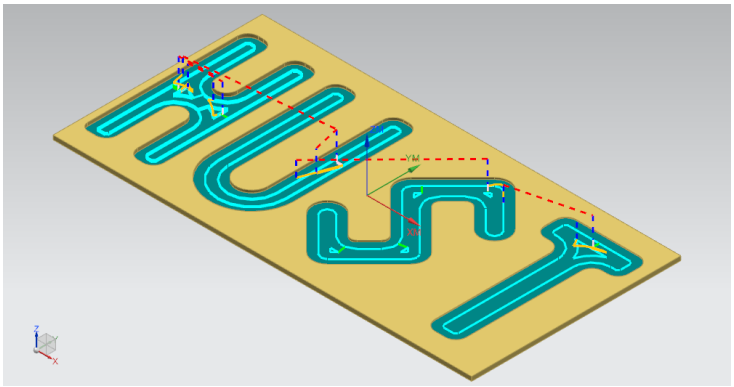


图 32 优化案例刀路

图 33 优化案例

应用加工过程主轴功率预测模型以及所提取的工艺参数，预测对应的主轴功率，然后利用多目标优化算法对 G 代码进行优化，应用所生成的 G 代码进行实际加工，可得优化后实际功率及实际进给速度。优化前后进给速度及优化前后实际功率对比如图 34、35 所示。对所采集的实验数据进行分析计算可得，优化后加工时间 78.54s，优化后主轴功率方差为 795。原始 G 代码加工情况下的加工时间为 89.46s，主轴功率方差为 986。因此加工效率提升 12.21%，负载波动降低 19.37%。

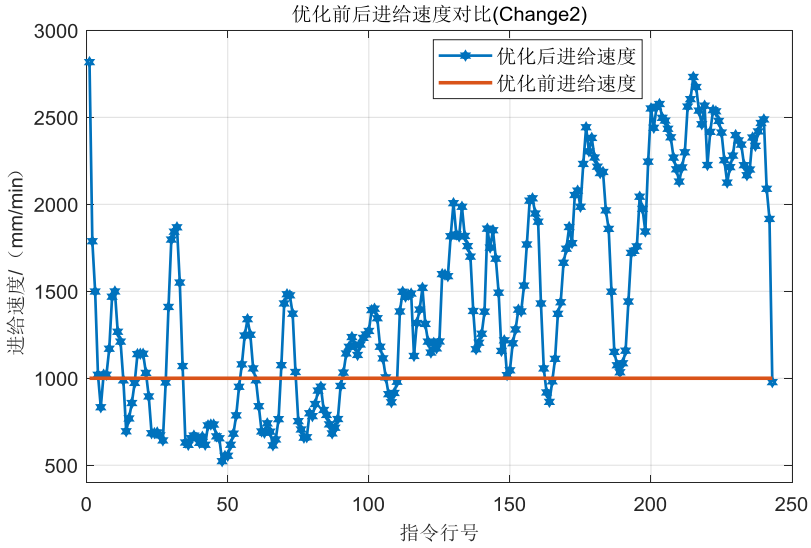


图 34 优化前后 G 指令进给速度

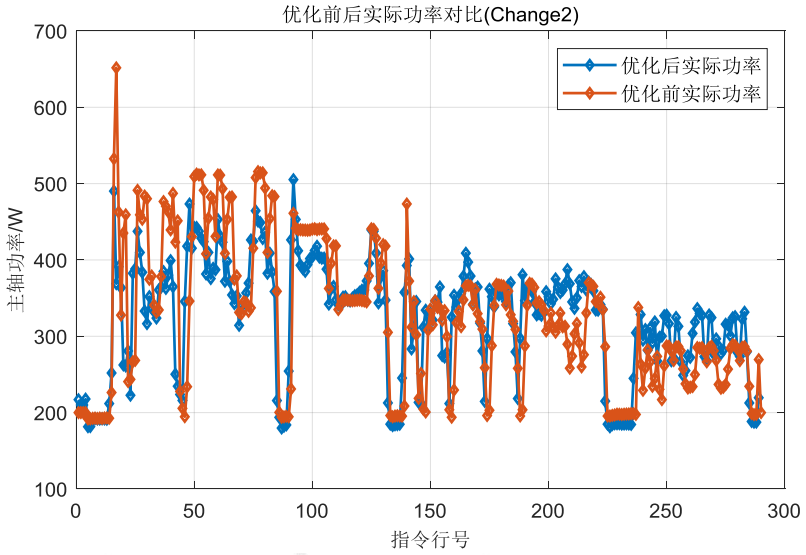


图 35 优化前后实际功率对比

3、总结

综上所述，本案例说明基于数控加工大数据和机器学习方法，可实现加工效率的显著提升，并为数控加工涉及到的诸如进给系统跟随误差预测、健康保障、生产管理等方面智能化技术与应用的实现提供了较好的参考。

三、案例 3：热轧带钢性能预报—清华大学、上海优也

（一）案例背景

本案例是我们对过去从事的数据建模实践工作的提炼和总结。我们用这个案例来说明，如何从工业大数据中获得相对准确、可靠的规律性结论。案例关注的重点不是具体模型，而是工业现场中遇到的某些一般性问题和处理这些问题的思路。

钢铁是最基础的工业原材料，被称为“工业的骨骼”。钢铁被广泛地应用在各种不同的场合，随着应用场景的不同，对钢铁材料的属性和质量要求不一样，为了适合不同用户的需求，人们开发出了不同品种和牌号的钢材。

钢铁产品的属性是由成分和加工工艺决定的。钢铁除了含有铁元素，一般还包含若干杂质和合金成分，这些杂质和合金是区分钢种最主要的指标，常见的杂质和合金有十几种。热轧是生产钢板(或称为带钢)的工序，热轧带钢是最常见的一种钢铁材料，约占我国钢铁产量的一半，在某些热轧厂，生产过的成分体系就有上千种之多。

力学性能是钢铁材料最重要的质量指标之一。用户订货时，往往对力学性能有特定的要求。力学性能也有很多种，如强度、塑性、韧性、硬度等，但指标之间往往存在关联性，这些性能一般通过标准试验来获得。例如，拉伸试验可以获得材料的抗拉强度、屈服强度和延伸率；冲击试验可以获得材料的冲击功等。事实上，用户最常见的要求就是针对拉伸和冲击所对应的性能。

钢铁冶金是一门相对成熟的工程学科，冶金和材料机理相对

清楚。在材料冶金领域，人们公认的原理是：材料的成分和加工工艺决定材料的内部微观组织，而（内部）组织决定性能。

但是，“组织”的概念复杂甚至抽象。首先，“组织”有不同的类型，被称为“相”，如马氏体、珠光体、析出相等，但部分材料由多种不同的“相”组成，这样就有了各种相的比例关系问题。一种“相”又需要若干相关的参数来描述，如位错密度、析出物大小、珠光体层间距等。组织参数往往针对显微镜下某个微小局部的统计结果，定量化的程度并不理想，例如显微镜下可以看到很多晶粒，但观察到的每个晶粒的大小都不一样，测量晶粒度时，人们往往只能对照标准图谱，给出级别指标，不仅精度低，主观性也强。

组织的形成是一个动态的过程，与成分、夹杂物和生产加工过程的参数相关。人们虽然有一些研究，但由于影响因素多、演变过程相当复杂，大量研究结果是在特定成分和工艺条件下数据拟合的结果，没有统一的理论公式。

很多人研究组织和性能之间的关系。以强度为例，人们发现了四种强化机制：相变强化、析出强化、固溶强化、细晶强化等，针对这些强化机制，有些理论或经验公式。Hall-Petch 公式最典型、最著名的理论公式之一，这个公式认为，细晶强化对屈服强度的贡献与晶粒度大小的平方根成正比，但即便是这种相对成熟的理论，也存在这明显的不足。比如，Hall-Petch 公式的系数本身与多种因素有关，而对系数本身的研究缺乏一般性、有深度的研究。

总之，相关机理知识往往是定性的、碎片化的、不精确的，这些理论成果难以拼出一个完整的模型。由于这个原因，钢铁的产品设计和研发往往离不开实验，实验不仅包括实验室实验，还包含大生产实验，一次大生产实验就可能需要数百吨钢水，时间和资金成本都很高。

上世纪 50 年代，英国学者 Irvine 等人提出了用数学模型预测钢材组织和力学性能的想法。人们普遍认为：研发这种模型的用途很广、意义很大。比如，可以用来设计新钢种、优化成分和工艺参数、推动一钢多用、满足用户个性化需求、生产组织优化、用预报结果代替检验等等。

上世纪 90 年代，基于数学模型的热轧带钢性能预报研究成为了世界性的热点课题。美国、加拿大以及德国等欧洲国家都专门投资研究。我国的宝钢研究院（技术中心）、钢铁研究总院、东北大学等国内院校，同时也开始关注这个方向，并进行了一些探索性的研究。其中，宝钢研究院先后专门多次立项进行了深入研究，前后长达十多年，直到取得相对满意的结果。

事实上，早在本世纪初，国外部分企业已有商业化产品出现。如西门子公司的 BM_MM 等、INTEG 公司的 HSMM 等。许多相关的报道都宣称自己的模型精度很高，用户却普遍不太满意。我们认识到出现这种问题的本质在于把用户需求简单地理解为模型“精度高”是不准确的。事实上，对用户来说，“模型精度”是个复杂的概念，而技术提供方把这个概念简单化了。成分和工艺不同，模型的精度不一样；而模型的用途不同，对模型精度的

要求和建模难度也是不一样。

很多产品自称模型精度很高，却不能满足用户需求。比如，有的模型平均精度高，却只适合简单的钢种；这些钢种的性能往往相对稳定，是不需要预报的。再比如，人们希望模型找出不合格的产品，但现实却是模型对合格产品的预测精度很高，对不合格产品的预测精度却低。模型用于新钢种设计时，往往缺乏训练样本，精度评估的难度也大。

我们认为模型精度应该是个综合性的概念，应该与钢种和使用场景结合，兼顾到适用范围、可靠性、可理解和可解释性。但在现实的条件下，特别理想模型是不存在的，我们必须针对具体的业务需求和条件建立适用的模型。

（二）解决方案

1、业务理解

从业务角度看，人们对模型的需求是多方面的。最常见的需求是用预报结果代替取样，以降低成本，加快交货速度。钢铁产品出厂时，钢厂需要给用户提供质保书，质保书上需要有力学性能的测量结果，检验过程要等到轧钢结束 2 天以后，还要剥掉钢卷外圈的一层，这项工作既延长了交货时间，又会导致材料的浪费。于是，人们希望用预报结果代替实际的检测值，人们一般允许预测结果与实际值有一定的差异，关键是正确判断是否合格，如果力学性能不能达到要求，就只能降级或者报废。

但是，完全用预报结果代替取样的想法并不现实。钢厂给高端客户供货时会受到质量贯标的限制，贯标要求必须取样，完全

用预报代替取样是不允许的，所以正确应用模型的方法是应与贯标允许的取样制度相结合，在保证质量和贯标要求的前提下减少取样。

模型的另外一个用途是钢种的优化。最典型的做法就是通过成分和工艺的优化，提高力学性能的合格率、降低合金成本或工艺成本。对于这样的工作，产品技术人员往往是有一定想法的，但吃不很准。如果直接通过大生产调整，存在着一定的风险，所以他们的需求是希望模型给出建议，以增强他们的信心、减少优化带来的不确定性。

模型的第三个用途是生产或质量管理。先进的钢铁企业往往是采用订单驱动、定制化生产的，每个用户的需求都不一样，这会导致产品的种类过多，进而导致生产组织困难、库存过多、材料使用效率低等。对此，钢铁企业往往致力于推进“大规模定制”，具体做法是在炼钢环节尽量把不同的合同归并到成分相同的钢种，而在轧钢环节采用不同的工艺、以满足不同用户的要求。利用性能预报模型，可以帮助人们选出适当的成分和工艺，模型还可以用于质量管理的智能化，比如当上工序的参数（如钢水成分）偏离理想区间时，可以动态设定下工序的工艺参数，避免产生不合格的产品，这在业内被称为“性能动态控制”。

模型的最重要的用途是钢种优化和新钢种设计。钢铁企业的产品开发，往往是通过实验室实验和大生产试验确定最终成分和工艺的，这个过程既费时间又费钱。大生产实验时，一炉钢水数百吨，一次实验的成本可能高达上百万元，有些钢种先后要试验

几十炉钢水才能最终定型，在这些实验过程中，有些是为了满足质量要求，有些是为了降低成本，但是每次调整和优化都是会伴随风险的，由于这个原因许多钢种的成分和工艺并不是最优的，只要用户接受就把成分和工艺固化下来、形成标准，不再优化了，但这些标准有不合理性，有的吨钢成本优化空间达上千元，有的则不合格率超过10个百分点，人们希望用模型解决这些问题。

每一种业务对模型的要求是不一样、建模的条件也不相同，最典型的差异有两种：

第一种是减少取样。模型应用针对的是大生产的钢种，有大量的生产数据，有条件通过统计方法建立模型。在这种场景下，模型误差大往往是数据采集异常导致的。

第二种是新钢种设计。在这种场景下，人们对模型精度的要求主要针对统计意义上的偏差，而不是特定的带钢，比如预报的结果与钢种实际性能的均值相差多少，这时模型针对的是从未生产过的钢种，所以最大的麻烦是没有直接的样本数据用于建模。从某种意义上说，使用模型的范畴超出建立模型的样本分布范畴。

模型其他的应用场景（如钢种优化、性能动态控制等），需求和特点往往介于上述两种场景之间。比如，钢种优化针对相对成熟的钢种，但优化范围可能部分超越原来钢种成分范围。

总之，人们对模型的要求是多方面的，既对精度有要求，也对适用范围有要求。单就误差来说，对具体预报的误差有要求，对平均误差也有要求，这些要求之间可能会互相矛盾，一般来说，适用范围小的时候，容易取得较高的模型精度。建模的目的不同，

采用的手段也就不一样，而这样的差别很多研究者并没有意识到。

人们往往想当然地认为：建模应该先从追求精度开始，精度足够高了，一切问题都解决了。还会想当然地认为：建模应该从一个成分体系简单、性能稳定的钢种开始，逐渐拓展到其他钢种。但现实却是：针对特定钢种建模时，各种关键因素的影响往往可以用线性模型逼近，线性模型的精度就接近极限了，采用高级复杂模型的价值不大，模型的精度往往与人的经验水平接近，但人们往往并不满足这种精度。人们还发现：模型本身的稳定性很差：针对类似的钢种建模时，模型参数差距非常大。事实上，即便是同一个钢种，采用不同时期的数据建模时，模型参数的差距也很大，这让人们对模型的可靠度产生怀疑。

导致这种现象的本质原因是数据本身的质量差。数据质量差时，模型的精度就无法提高。提升数据质量的代价是非常大的，效果却不一定明显，为此模型评估成为一个必须认真研究的新问题，其相关的研究也先后持续了 10 多年。

2、数据理解

从机理上讲，力学性能决定于材料的组织，组织决定于成分和工艺，数据理解要从“性能”、“组织”、“工艺”这三个方面入手。

力学性能其实有多种，最常见的是拉伸性能。拉伸性能一般包括抗拉强度、屈服强度、延伸率等几个指标。其中，区分强度有分为上屈服、下屈服等很多种类，测量结果是类似的，但却是有差异的。另外，有些客户还要求测量冲击性能，冲击性能主要

指不同温度下的冲击功。

“组织”的概念比较复杂。狭义的“组织”信息是显微镜下面观察得到的图像信息，图像信息中含有某些数字化的指标，如珠光体比例、层间距、晶粒度大小、析出物大小等。广义的组织其实是包含合金或者杂质元素的，这些元素是固溶、化合物、析出等形式存在的钢材中，对力学性能的影响很大。组织指标是显微镜视野中局部的统计特征，量化得不够精细，例如晶粒度是按级别区分的。在某些钢种的试样中，力学性能的测量值各自不同，但晶粒度的级别却不发生变化，所以晶粒度的测量结果难以用于准确的性能模型。

原则上讲“工艺”针对炼钢、连铸、轧钢等生产环节，但重点是热轧的加热、轧制、冷却过程。这个过程的工艺主要用温度、变形描述，而这些参数又都是时间的函数。遗憾的是，这些参数一般是无法连续测量的、甚至难以准确测量的。在具体实施过程中，考虑到检测和测量条件，一般用个别关键位置的温度和（中间坯、带钢）厚度来表示工艺。

3、数据准备

性能预报模型大体有两类：机理模型和统计模型。在行业内部，所谓的数据建模，指的是通过工艺和成分预报力学性能；机理建模则是通过工艺成分预报组织，在此基础上预报性能。由此可见，两类模型要求的数据是不一样的。

绝大多数用户只要求检验力学性能（主要是拉伸性能）而不要求测试组织。所以，企业中积累的数据主要与力学性能有关，

而缺乏组织方面的数据。受技术和经济条件的约束，工厂对工艺数据的记录和存储并不理想，比如控制温度往往只是记录带钢的平均温度，但带钢上的温度波动还是比较大的，力学性能的测量是从带钢的某个点上取的，即便有了整条带钢上的工艺参数，取样点位置与参数的对应也存在问题，于是一般只能用带钢的平均工艺参数值来代表取样位置的工艺参数，这当然会产生误差。

4、数据建模

（1）先期尝试

基于数据条件的考虑，人们一般用成分和工艺直接预报力学性能。

前人曾经尝试过各种模型，但模型精度不是很理想。事实上，简单的线性回归模型与各种精心研制的神经网络模型也没有显著差别，这是有道理的。针对一个钢种进行建模时，由于成分和工艺参数的变化相对较小，线性模型（或考虑个别非线性因素的可加模型）应该能够比较好地逼近实际情况，实际的研究结果也证明了这个猜测。于是，问题变成了：既然性能模型可以用线性模型来近似，模型精度为什么不能提高了？要回答这个问题，需要对数据本身进行反思。

（2）数据理解深入

如前所述，人们习惯于用“精度”来衡量模型，其实就是用平均精度来衡量，这种衡量方式似乎是天经地义的，但其实是有问题的。人们对精度的理解却并不深刻，这是导致很多研发团队陷入误区的深层原因。

用户关心的精度是模型使用时的精度，评价模型时只能用历史数据的平均精度。对于性能预报模型，两种精度的差异可能非常大，而且随着建模时间与模型使用时间间隔的拉大，两者的差别会有越拉越大的趋势。导致这种现象的本质原因是建模数据的分布特征与使用时不一样，即便是针对同一个钢种，不同年份的线性模型都会存在显著的差别。我们知道：在标准确定之后，同一个钢种的成分和工艺基本不变，为什么还会存在这种现象呢？

其实，标准都是相对的，会存在变化的空间，比如 C 含量范围是 0.15 到 0.18，Mn 含量从 0.4 到 0.6，但是标准并不会规范 C 和 Mn 的相关系数，而炼钢时所用的合金不同，两者之间的相关系数就不一样。另外，检测误差的大小及其相关性，也会发生变化的，所以原始数据的统计特征会随着时间变化，这些统计关系的变化，会影响模型的参数。所以，模型对历史数据的精度高，使用时的精度未必高。如果仅仅把减少模型平均误差作为优化指标，过拟合就可能非常严重。

即便是对于历史数据，模型精度也存在着不可逾越的极限。其实，当模型精度达到一定水平后，误差的主要来源可能不是模型本身的偏差，而是检测数据的误差。不难理解，即便模型完全正确，有误差的输入会导致有误差的输出。这个道理大家都懂，但却往往对误差的影响程度估计不足，很少有人意识到检测数据的误差是模型误差的主要来源。

输入数据的误差是可以估计的，同一个位置测量两次、同一个试样测试两次，结果也会存在差异，这就是典型的检测误差。

我们用这种办法，可以测算误差的标准差并记为σ_1，同时把同一个钢种中所有的检测结果都拿过来，可以计算相关参数的标准差，并将这个标准差记为σ_2，σ_1与σ_2的比值就表明某个环节检测误差的影响。我们发现，检测误差的影响是非常大的，比如在某些情况下，延伸率σ_1与σ_2的比值在 0.4~0.8 之间，C、Mn 等合金的比值也经常在 0.6~0.8 的水平。可以说，在一个钢种内部，大多数关键输入和输出变量的检测误差都是相当惊人的。这就意味着：如果把一个钢种作为对象建模，测量误差占的影响会非常大，模型精度不可能太高。

实践证明，把模型精度作为模型优化唯一的目标时，存在两个方面的危害。

首先是浪费研究时间。有人为了提高精度，花费数年乃至十多年的研究新的模型和优化算法，只比线性回归模型提高了几个百分点。

其次，会导致模型的系统性偏差。数据建模往往普遍采用误差最小化的方法，人们熟知的是建立神经元网络模型时，一味地减少误差，容易引起过拟合问题。其实，以误差最小化为目标时，很多模型都有类似问题。一元线性模型可谓最简单的数学模型，最小二乘法是最常见的优化方法，采用最小二乘法建立一元线性模型时，也存在问题，可以证明：当自变量存在检测误差（或者说检测误差难以忽略）时，线性回归是“有偏估计”，模型参数估计值的数学期望并不逼近真实值，用通俗的语言说就是误差最小的模型并不是正确的模型。所以，在特定条件下如果单纯追求

误差最小，则会扭曲客观规律，而正确的模型并非误差最小的那个。

另外，实践和理论推导都表明，由于检测误差相对较大，针对单个钢种建模时，要有 2000~20000 组数据，参数估计才能稳定下来，而样本量能够达到这种要求的钢种，其实非常少。

（3）易于评估的建模

为了应对数据误差问题，我们找到了一种评价方法，下面以厚度为例，介绍这种评估方法。

同一个钢种，往往会生产不同厚度的产品，厚度变化后，轧制的变形会发生变化、冷却强度也会发生变化，进而会引起性能的变化。人们往往会针对厚度的不同，调整冷却制度等工艺参数，以保证特定钢种的性能（主要是强度）基本稳定。

同一个钢种、同一个厚度级别下，可能会有很多样本。由于样本来自同一个钢种，成分和工艺目标值都保持不变，这个时候的参数波动，基本上都是围绕目标值的随机波动，波动本身可以近似服从正态分布,没有系统性干扰。这时，我们把同一个厚度级别中样本力学性能（如抗拉强度）取平均值，把厚度和性能的平均值呈现出来，就得到类似下面的这样一条曲线。

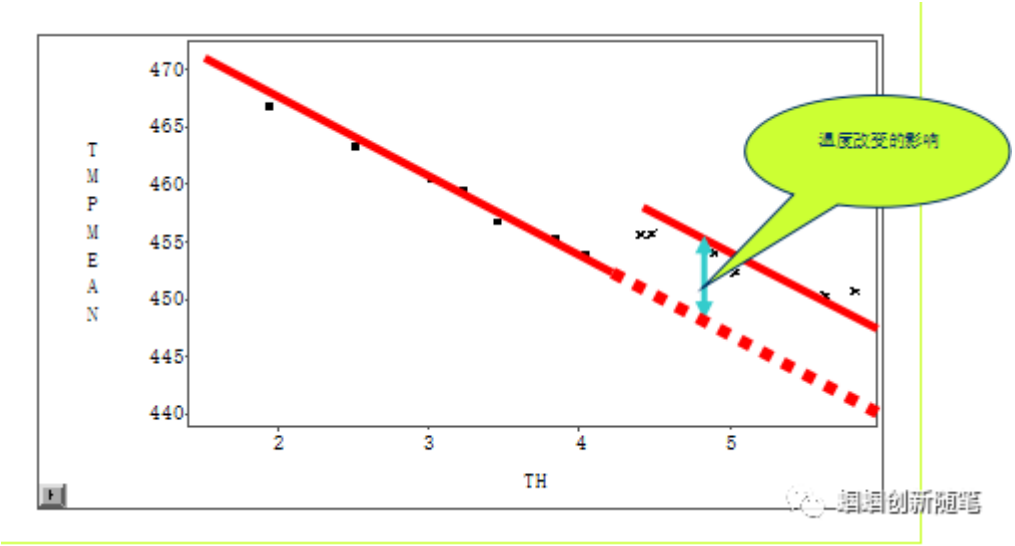


图 36 温度改变产生的影响

我们发现，除了一个跳跃点，力学性能指标和厚度呈现稳定的、近乎线性的连续关系，而在这个跳跃点，正好对应某个工艺参数目标值的改变。不难理解，直线关系本质上是厚度对性能的影响，而这个跳跃则代表着这个工艺参数的作用。

这条曲线可以用线性回归拟合出来，当样本足够多时，随机的干扰往往可以通过平均值滤除了，所以拟合出来的曲线是可信的，我们认为这种平均值之间的曲线反映了变量之间的本质联系。

这种建模方法既直观又符合工艺技术人员的习惯和冶金学原理，可信度极高，模型的适用范围一目了然。这种模型的意义本质是通过数据的处理、提升数据的精度，并得到唯象的结论，类似于天文学上的“开普勒定律”，而这样的做法，往往只有在“大数据”的背景下才能展开。

我们还可以用这种办法评估其他一些关键变量的影响。我们设想：如果所有重要的工艺参数，都能用这种办法评估出来，就

能把各个变量的影响搞清楚，然后再根据这些分析结果建模。

（4）建模的深化

上述建模的思路容易理解也易于验证，但是在实施过程中却不宜推广，这是因为厚度的检测误差是一个可以忽略的变量，而目标温度改变时，目标值的差异基本上等于实际值平均值的差异。

这样的数据条件，并不适合所有的变量。比如，在一个钢种内部，成分的目标值本身是不变的，成分的检测误差往往符合正态分布，而且误差不能忽略，采用这种办法时，就会产生显著的“有偏估计”，而且对于一个钢种，工艺参数目标值变化得并不多，有的钢种甚至没有变化，所以上述做法虽好却不具备一般性。

这时，需要进一步突破传统观念。一般人认为：模型研究应该从个别到一般，具体地说就是先把单个钢种的特性研究透，再研究跨钢种的问题。但事实上，由于前面讲到的原因，这条路是走不通的，表现为分析结果很不稳定、可靠度很低。比如，按照同样的逻辑分析，不同年份的分析结果是不一样的，不稳定就意味着对规律的认识不清晰。

所以，针对现实的数据情况，针对单个钢种的研究存在极大的困难，其本质原因就是数据的信噪比很低。由于单个钢种的成分和工艺往往集中在某个工作点附近，参数的波动和检测误差常常处于同一个级别上，所反映的规律就是扭曲的。

为了解决这个问题，可以把不同钢种的数据放在一起，进行建模，这样做的好处是跨越钢种以后成分的目标值经常发生改变，这时数据就分布在不同的工作点上，数据的信噪比会极大地提升。

但是，多钢种共同建模也会遇到困难，这个困难就是多变量、非线性。我们知道：在一个钢种内部，许多工艺和成分参数是相对稳定的，模型可以用线性模型来逼近，但是如果建模的范围跨越了钢种，这些假设条件就不再成立了。我们必须直面真正的非线性、多变量问题。

如果数据理想，多钢种建模的思路是这样的：

先对一类机理简单的钢种（普通碳钢）展开研究，这类钢种的成分体系比较简单、影响性能的工艺参数比较少，虽然是多变量对象建模，但重要变量的数量少，虽然是针对非线性对象建模，但非线性的特征弱。

根据我们对机理的认识，估计主要成分的作用是可加的或者是近似可加的。于是，在数据条件理想的情况下，可以分别研究各种成分的作用，具体的做法是：从海量的样本中找到若干个特殊的钢种，这些钢种之间只有一个成分有显著的变化，其他关键变量保持不变，这样就可以比较精确地估计这个成分的影响了。

然而，现实中却找不出这么多分布理想的钢种，钢种变化时往往会有多个成分或者工艺参数的目标值发生变化，为了解决这个问题，我们采取“逐次逼近方法”：

假如有两个因素 X、Y 同时提升了强度，我们先选取样本集合，找 Y 变化较小、X 变化较大的钢种组，按照前面的办法分析 X 因素的作用，X 的影响搞清楚后，就可以在模型中扣除 X 的影响，得到一个“扣除 X 影响的、虚拟的强度指标”。不难理解，X 对“扣除 X 影响的虚拟强度指标”的影响变得很小了，甚至可

以忽略不计。

我们假设 X 的这种作用，适合特定范围内所有的钢种，我们在这种条件下，可以考虑 Y 因素对“扣除 X 影响的虚拟性能指标”的影响，这时重新选择一批 X 变化相对较小、Y 变化较大的钢种数据，研究 Y 对“扣除 X 影响的虚拟性能指标”的影响。

在此基础上，可以循环迭代求精。比如，研究 Y 对“扣除 X 影响的虚拟性能指标”的影响搞清楚之后，可以把原始的性能中扣除 Y 的影响，得到新的“扣除 Y 的虚拟性能指标”，再用这样的指标重新组织数据，去研究 X 的作用。经过若干次迭代，各个因素的影响就会越来越精确了。

可以证明，如果 X、Y 的影响是线性的，这种办法与线性回归的结果是一样的，但这种做法却适合 X、Y 的影响各自是非线性的情况。这样做的前提是两个变量之间没有交互作用，所以做这种研究的时候，尽量根据机理，首选那些理论上交互作用弱、数据条件好的来做，做完之后还可以用现实的数据进行验证，如果确实发现过程不收敛，也容易发现并研究交互作用的表示方法。

事实上，单个要素的影响搞清楚之后，要素之间存在的“弱非线性”也就容易发现了，可加模型的误差，就反映了这种非线性因素。实施上，这种非线性因素也是机理上可以解释的，经过长时间的研究，我们用了一个非常简单的模型，描述了数百个钢种、上千种工艺组合的模型。从某种意义上说，对于普通碳钢，人们已经无法找到更加简单、参数更少的模型了。

在此基础上，可以扩大钢种的范围，加入新的成分后，产生

新的强化机制，上述方法同样可以用于研究其他强化机制的作用。

在成分体系扩大化的过程中，会遇到多种元素形成化合物或者沉淀物的过程，在这个过程中化学元素不是单独起作用的。我们根据化学原理建立机理模型，计算化合物的生成，并把这些计算值作为输入建立模型。根据机理，化合物对力学性能的作用，往往是线性可加的，所以这样做可以在重复利用冶金机理的前提下，简化数据建模。

理论上讲，本文提出的做法并不能保证成功，但现实却是幸运的。这样，预测模型可以覆盖90%以上的热轧钢种，涉及到近千个成分体系、数万种工艺和规格的组合。

（5）数据认识的再深化与模型求精

采用跨钢种建模策略以后，会发现有些难以解释的误差难以消除，我们意识到这是因为还有些系统性的误差在起作用，进一步研究发现很多系统误差是测量过程导致的。比如，同样是“屈服强度”，可以有“上屈服”、“下屈服”等若干种类，在取样的过程中，有横向、纵向等差别，取样方式有冷态和热态的差别、取样的位置和方法也不一样，温度测量有黑度系数的差别等等。所以，必须明确的一点是“测量结果”的系统误差是很显著的，这些系统误差必须给予补偿，才能把误差消除。

事实上，模型精确到一定程度以后，系统性干扰的影响就会变为主要矛盾，需要专门建立若干子模型模型，评估这些因素的影响。事实上，我们的建模方法往往是从重要的变量开始考虑，当重要变量的影响搞得比较精确以后，次要变量的影响也就很容

易凸显出来了。这样，模型的可预测范围也就逐步扩大了，囊括了大多数的热轧带钢。进一步的分析还表明将模型适合用于不同的（传统）热轧产线时，模型的参数几乎不需要调整，这也表明模型参数反映的是冶金学的客观规律。

随着模型的本体精度的逐步提高，很多小样本、试验性钢种的预报精度也逐渐提升了。我们还从这些小样本数据中，发现了一些过去人们不清楚的冶金机理。从某种意义上说，大数据给我们一个新的方法来研究科学问题。

在可见的文献中，性能预报都是针对拉伸性能的，但是越来越多的高档产品希望对冲击性能有具体的要求。在测量过程中，冲击性能是一种波动性较大的性能，若利用前面建模过程形成的经验和方法，能建立起精度相当高的预测模型，这在世界范围内都是具有突破性意义。

如前所述，钢种不同，针对具体样本的预报误差的标准差并不一样。我们分析了这种现象的原因：每种成分的控制和测量误差是不一样的；成分体系不同，对工艺的敏感度不同；钢材的组织不同，力学性能测量的误差也不一样。于是，我们设想：预报误差本身是不是可以预报的？

可以预报误差的前提是模型的误差主要不是模型本身的参数导致的，而是各种检测误差和随机不可见因素导致的，而误差的预报，并不是预报具体样本的误差，而是误差分布的特征，如标准差。

于是，我们可以建立误差模型，其输出是预报误差的标准差，

输入是成分、工艺和测量方式。通过数据建模，确实印证了前面的猜想：预报误差是可以预测的。

这样，给定成分、工艺和性能要求后，就能预报产品的合格率。由于模型的适用范围很广，这种模型可以广泛地用于钢种的设计和优化。

5、模型验证

如前所述，在数据质量相对较差的情况下，模型误差不可能无限地逼近与零误差。更严重的问题是：在误差接近的前提下，模型参数可以差距很远。如果我们假设模型是符合客观规律的，这种现象就不应该出现。可以肯定的是：如果允许这种现象出现，当预报结果超出建模数据范围时，误差就会急剧增大，这样的模型既不适合新钢种设计，也不适合钢种的优化、材料集约使用和动态控制。

为了解决这个问题，我们采取了一种“平均值对平均值”的验证方法，具体地说就是把同钢种、同工艺的力学性能预报结果取平均值。这样，当组内的样本数足够大时，随机误差就会被平均掉，然后用预测的平均值去预测实际取样的平均值，而两者的差别大体能体现模型的预报误差。

模型的可靠性取得还基于两个原则：一个原则是符合冶金机理，另一个原则是模型尽量简单并便于理解。模型总体上由多个子模型过程的，除了化合物和沉淀物计算采用理论方法外，几乎都是由简单的线性模型、一元模型构成的，只有个别因素之间存在非线性关系。子模型不仅简单，而且在所有的钢种中保持不变，

这样模型的可靠度和适用范围就大大提升了。

（1）大范围验证

模型能对绝大多数的热轧产品进行预报。这些钢种的屈服强度从 160MPA 到接近 800MPA，工艺和成分体系的变动也极大，比如 C 含量从 0.001%到 0.7%，Si 含量从 0.001%到 3%，除了极少数包含特殊合金和强化机智的钢种外，几乎都可以准确预报。

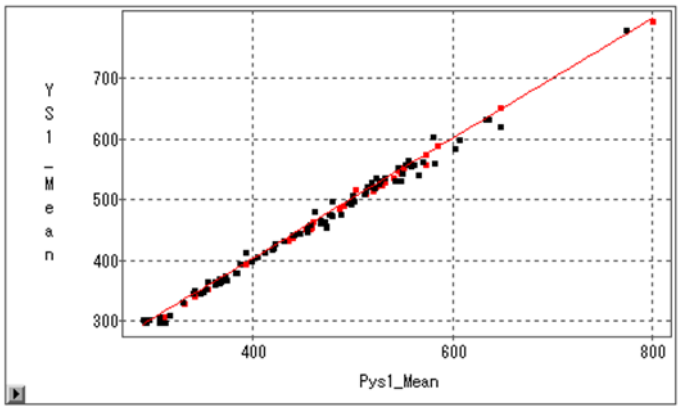


图 37 特征相关系数

（2）小样本钢种验证

根据前面的分析，模型参数对各种钢种都是一致的。在建模、确定参数时，样本量大的场景影响更大，有数以百计的钢种的样本数量少、对建模参数的影响几乎可以忽略不计，但对他们的验证结果仍然是相当满意的，和大样本钢种的预测精度是非常接近的。

（3）特殊钢种验证

我们选取了一些建模时尚未生产的钢种以及由于特殊原因未能纳入模型体系的钢种进行验证，验证结果都是非常满意的。比如，我们建立模型时，Si 的成分主要在 0.5%以下，最多不超

过 1.5%，我们用 Si 含量 3%以上的钢种测试模型，多个样本的误差不足 1%。

（4）相关性验证

单个钢种的参数波动很小，应该可以用线性模型逼近，但是由于干扰和相关性等原因，线性模型的系数很不稳定，导致模型的外延性很差，无法用于钢种的优化与设计。

我们采取了一种特殊的方法，印证了模型的有效性。对于每个钢种，我们建立了两类模型：一类是非线性的全局型模型，一类是针对特定钢种建立的线性回归模型。在此基础上，比较两种预测结果的相关系数。结果发现，几乎都对所有的钢种，两者的相关系数非常高，许多相关系数高达 0.95 以上，两个预报结果的一致性，说明全局模型的预测与局部表现一致。

这就印证了我们的观点，单个钢种可以用线性模型逼近，模型误差和系数的差异，主要是检测误差和数据分布关系导致的。

（5）实验验证

有些钢种生产不稳定，我们按照模型的计算，给出改进方案，结果完全达到了模型预定的结果。

我们还发现，某些理想的做法，是大生产无法模拟的，于是我们采用实验室验证，结果表明模型的预测完全是正确的。

（6）异常验证

针对预报误差特别大的样本，我们进行分析后发现有两种情况：

混钢水。某钢种中，我们找出所有预报误差大于 25MPA（约

为强调的 10%）进行分析，发现这些样本中 92%有混钢水标志，也就是带钢的某些部分有两种不同的钢混合。按照惯例制度，带钢的成分是带钢所属炉次的成分，带有这种标志的带钢，试样上的实际成分与标定成分可能有很大的差异，实际样本分析也印证了这个判断。

检验误差。部分误差大的带钢同时也出现了力学性能不合格的现象，当出现力学性能不合格时，按照规定需要重新取样测试，对重新取样测试的试样进行分析，发现模型误差显著减少、归于正常。

（7）对非线性问题的说明

总整体上看，模型是非线性的，但事实上模型是由若干简单的子模型叠加而成，非线性体现在子模型中，除了个别用基本科学原理计算的子模型外，用数据分析获得的子模型都非常简单，有的是单变量的非线性函数，多变量的非线性模型也都很简单，如双线性模型。

模型简单的好处是容易验证。科学哲学认为，当模型精度足够高、无法更简单的时候，就可以认为模型是科学的。前面曾经提到：我们验证子模型的办法，是首先扣除其他因素的影响，再去除随机干扰，这样处理后线性相关系数经常可以达到 0.99，而同一个子模型要用于数以百计的钢种，我们没有理由不相信模型的正确性。

但是，做到这一点是需要有前提的：其他子模型必须准确、数据选择和处理要合理。为了具备这样的前提，往往要花费大量

的时间——时间是以十年为单位计算的。

（三）实施效果

正如人们所期待的，性能预报技术可以用于解决多类问题。

1、精准选样系统

如前所述，性能预报模型最为人熟知的应用是代替取样，但是这个功能的实现与用户和贯标要求有关，通过选择某个合适的场景，根据预报结果决定是否取样。该项目投运后，每年减少取样 60%以上，节约成本 150 万元左右。

2、钢种优化

有家钢铁企业对 20 多个传统钢种的成分体系进行了优化，每年取得 2000 万左右的经济效益。

有一类钢种叫做包晶碳钢，这种钢在浇铸过程中容易形成缺陷，需要进行处理，每吨钢的处理成本 30~50 元。解决问题的办法，是重新改变成分体系，避开这类钢种，但对成分的改变可能会改变力学性能。对此，材料专家根据模型的计算，确认新成分体系，在性能符合要求的前提下避开包晶碳钢的区域。

S 是钢水中常见的杂质，为了降低 S 含量往往要经过 RH 精炼工序，这样每吨钢会增加几十元的成本，而对某些用户，S 含量没有必要脱得太低，不需要经过 RH 精炼，但工厂害怕这样做会影响力学性能，不敢轻易取消 RH。经过模型计算后，认为不经过 RH 同样会满足用户需求，实践证明这种做法是可行的。

3、集约化生产

有两种钢种的成分体系类似，人们想把两种成分体系合二为

一，但却发现两者的力学性能有很大不同。经过模型计算发现，力学性能的不同，是取样方式的差别导致的，而不是成分和工艺导致的，后来将二者合并，未出现任何意外。

4、新钢种设计

有一种高等级钢种的力学性能一直不稳定、强度经常达不到要求。为了解决这个问题，工厂在每吨钢中加入某种合金，让吨钢成本增加了约 1000 元。后经模型测算，只要将加热温度提升 20 度，就可以起到更好的效果。

某企业某新产品合格率较低，一直无法有效地解决，无法批量供货。经过模型测算后发现，该厂的取样方式不规范，导致了性能不合，改变取样方法后，合格率就显著提升了。

四、案例 4：智能化橡胶浆液浓度控制—京博石化

（一）案例背景

橡胶行业是国民经济的重要基础产业之一，它不仅为人们提供日常生活不可或缺的日用、医用等轻工橡胶产品，而且向采掘、交通、建筑、机械、电子等重工业和新兴产业提供各种橡胶制生产设备或橡胶部件。其中，丁基橡胶（IIR）由于突出的气密性和水密性，良好的化学稳定性和热稳定性，广泛用于生产机动车轮胎内胎、密封材料、电绝缘材料、防毒面具及医用瓶塞制造方面等。由普通丁基橡胶与卤化剂发生氯化或溴化反应可以得到卤化丁基橡胶（HIIR），HIIR 不仅保持了 IIR 优良的特性，同时具有更好的物理强度、减振性能、低渗透性、耐老化等特性，应用前景广阔。

目前，卤化丁基橡胶的生产方法主要有干法和湿法 2 种。干法又称干混卤化法，是将成品丁基橡胶和卤化剂通过螺杆挤压机，在机械剪切作用下，对丁基橡胶进行卤化，生成卤化丁基橡胶。湿法又称溶液法，是丁基橡胶与卤化剂在溶液中进行反应生成卤化丁基橡胶，最常用的溶剂为正己烷。国外丁基橡胶装置生产的产品 85%以上为卤化丁基橡胶，尤其是埃克森、朗盛等国际巨头在制备工艺、生产设备、操作管理等方面具有巨大的优势。国内丁基橡胶装置生产的多为普通丁基橡胶产品，与国外还存在较大的差距，主要表现为:

1） 丁基橡胶的原料是异丁烯和异戊二烯。这两种原料主要来自于乙烯裂解的C_4、C_5馏分。国内的C_4分离技术和具有自主

知识产权的 MTBE 裂解制聚合级异丁烯生产技术与国外还存在明显差距。

2） 丁基橡胶溶液浓度配比缺乏有效控制手段。尤其是在湿法溶解丁基橡胶的过程中，缺乏有效地浓度控制手段，导致后续卤化工艺控制困难，产品质量无法得到保障。

3） 国内丁基橡胶产品牌号较少、用途相对单一，新产品研制和开发力度不足，专业技术人才稀缺，生产工艺与设备相对落后，在产能、能耗、物耗等方面控制还有着较大的提升空间。

在卤化丁基橡胶生产过程中，丁基橡胶胶粒水溶液浓度的有效控制是保证生产工况稳定、安全生产以及高质量生产的关键。但是，由于工艺设计以及生产设备的局限，国内大部分湿法制备装置中胶粒水溶液的浓度无法用传感器直接测量，大都靠人工经验来判断，由于工艺链路复杂、过程持续时间长，技术人员的判断存在误差等因素，经常导致实际生产过程中长期处于波动状态，难以稳定，直接影响产品质量的控制。

同时在卤化反应阶段，液相反应釜中浆液浓度控制也是至关重要的，反应釜中需要有稳定、适当浓度的橡胶浆液参与卤化反应。在工艺流程与生产设备不变的情况下，浆液浓度适当增加，装置负荷可控的前提下可以有效降低产品的单位能耗和物耗。如果浆液浓度过高，浆液传热速率下降，聚合反应热增加，极容易引起结块，甚至发生“暴聚”事故。如果浆液浓度过低，装置负荷效率下降，反应不充分，影响装置产能。因此，为保障装置的高负荷、长周期稳定生产，亟需解决胶粒水溶液的浓度测量问题，

对卤化反应阶段的胶浆浓度进行有效控制，提升装置的综合运行效能。

本案例基于工艺流程设计、生产过程数据、设备运行机理等多维信息，利用操作经验、运行机理与数据科学融合建模方法，实现了胶粒水溶液的浓度预测，从而为后续的胶液配比、卤化反应、产品质量控制等环境提供基础保障，有利于改善湿法制 HIIR 的生产水平，助力我国工业高质量发展。

（二）解决方案

1、业务理解

（1）技术路线

本案例以连续溶液法制备 BIIR 的生产装置为研究对象，针对丁基橡胶胶粒溶液浓度开展“软测量”方法研究，基于过程工艺参数、设备运行数据、产品质量化验数据等多维度信息，分别建立橡胶胶粒浓度预测模型与机器视觉模型并融合，实现胶粒水溶液胶质含量实时预测，从而有效指导操作工对聚合反应前的胶浆液配比操作，保障装置的安全、稳定、高效生产。主要研究内容包括如下几个方面：

1） 针对普通丁基橡胶、卤化丁基橡胶的制备过程进行研究，根据连续溶液生产工艺特点，重点对生产过程中的工艺参数、关键设备运行机理、人工配比操作经验、产品质量影响因素等方面进行分析，结合工业机理与操作经验，开展有效样本收集工作。

2） 对装置生产过程中产生的工艺数据、设备数据、化验数据等多元信息进行分析，面向工业时序与关系型数据的特性研

究，开展特征选择、转换、组合与提取等特征工程工作，识别橡胶胶粒浓度“软测量”关键影响因子。

3） 建立橡胶胶粒浓度预测模型，选用 BP 神经网络算法，结合特征工程识别关键因子与有效样本集，开展模型训练与评估。

4） 建立机器视觉模型，为后续大数据预测模型提供更多数据支撑，从而提升整体算法的可靠性与鲁棒性。

5） 选取应用验证环境，将橡胶胶粒浓度预测模型进行部署，根据真实工况运行数据，分别评测模型在稳态运行工况以及异常扰动工况下的模型性能，从而对模型进一步调优、上线投运，实现橡胶胶粒浓度的实时预测，为操作工提供胶浆溶液配比建议。

从卤化丁基橡胶生产特定场景出发，剖析卤化合成过程波动现象，提出胶浆手工配比干扰问题。结合装置生产过程中的实时工艺参数、设备运行机理、人工操作经验等多维信息，提取关键特征、建立胶粒浓度预测模型、机器视觉算法，将训练与评估后的模型在实际装置上进行应用验证，进一步对模型在不同生产工况下的在线预测性能进行调优与评价，最后对研究方法进行总结梳理，形成可复制、可推广的工业智能场景应用创新方法，面向不同的卤化丁基橡胶工厂以及不同行业、不同领域进行推广。

本案例的技术路线如下图所示：

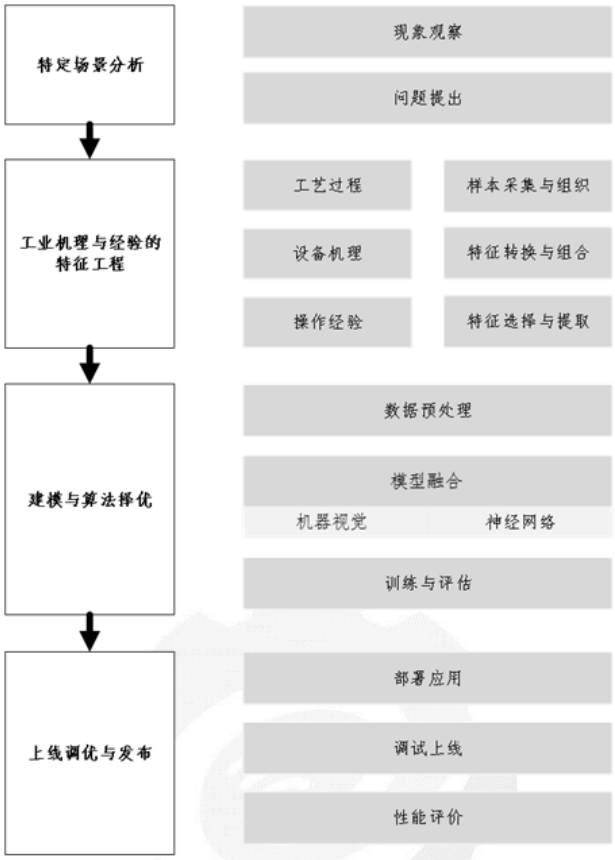


图 38 智能化橡胶胶粒浓度控制技术路线图

（2）工艺介绍

本文以京博石化某橡胶车间为试验对象，其生产过程中采用 DCS 控制系统对丁基橡胶的工艺流程进行控制，橡胶溶液配制所使用的主要原料是正已烷和橡胶原液。首先将橡胶的水悬浮液经过挤压机滤去大部分水，然后与输送来的热正已烷在 A 罐进行混合搅拌，再通过输送泵送到 B 罐，与输送来的冷正已烷进行混合搅拌，再输送至 C 罐，与输送来的冲洗正已烷进行混合搅拌，经过三级配制形成浓度约 15%的橡胶溶液，最后输送至 D 罐进行存储，同时也在 D 罐进行化验采样，用以反馈橡胶浆液浓度控制。

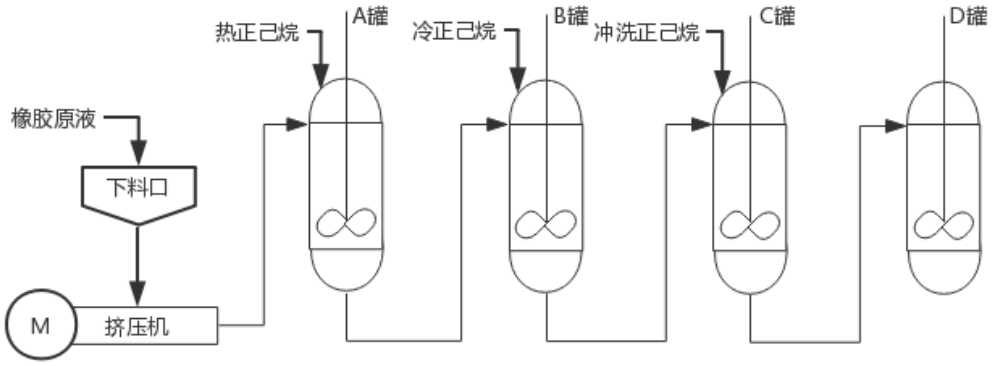


图 39 橡胶浆液配比示意图

浆液浓度的定义如下:

$$\text{浆液浓度}\ (\mathrm{g}/L)=\frac{\text{橡胶原液颗粒流量}\ (\mathrm{g}/h)}{\text{橡胶原液流量}\ (\mathrm{g}/h)+\text{热己烷流量}\ (\mathrm{g}/h)+\text{冷己烷流量}\ (\mathrm{g}/h)+\text{冲洗己烷流量}\ (\mathrm{g}/h)}$$

正如上图中显示，胶浓配比过程中通过引入 A、B、C 三个缓冲罐来稳定浆液浓度。其中，A 罐添加的为热己烷，主要目的是使罐内溶液温度上升，橡胶干粒充分溶于己烷溶液中；B 罐添加的为冷己烷，主要目的是使罐内橡胶浆液温度降低稳定至 55-60℃，从而满足后续卤化反应生产温度要求；C 罐添加的为冲洗己烷，主要目的是防止已经生成的橡胶浆液凝固在罐壁内。因此，C 罐的冲洗己烷量一般情况下不会做任何调整，基本稳定在 3600kg/h，而前面两罐中热己烷和冷己烷一般是对半等量的，稳定在 17000kg/h 左右，但是当 B 罐温度过低时，就会多添加一些热己烷，少添加一些冷己烷，当温度过高时，反之即可。

目前，车间采用四班倒的工作制度，每班会在工作薄上记录橡胶浓度化验数据，工作薄因此也是交接的主要内容。

尽管工艺上主要是依据胶浓化验数据来调节各环节的己烷添加量，但是每个班次有不同的调整经验，所以效果也不尽相同。

比如当出现较大偏差时，有的班次会以较大的调整量进行调整，使浓度可能一次性恢复正常，也有可能造成反向波动，有的班次会逐步调整，步步逼近至稳定合适的胶浓。当出现较小偏差时，有的班次会进行一些已烷量的微调，有的班次则不采取任何措施等待其进一步变化。由于不同班次操作人员的操作经验不同。

2、数据理解

橡胶生产线属于典型的流程与离散特征结合的生产过程，其中丁基橡胶合成、分离、卤（溴）化反应属于典型的流程过程，涉及多种静设备（如聚合反应器、热交换器、汽提塔、罐）、动设备（如泵、搅拌器）、仪表设备（阀门、控制系统），过程中的温度、压力、流量等参数通过 DCS 系统进行管理；胶粒水溶液脱水、胶浆配比、干燥、压块、装箱等环节属于典型的离散过程，涉及多种成套设备（如振动筛、挤压机、膨胀干燥机、压块机）与滚筒流水线，过程中的设备运行参数、物料温度、工序状态等信息通过 PLC 系统进行管理。同时，生产过程中的用电数据（如电流、电压、用电量）通过厂内变电站的综保系统进行管理。

一般橡胶生产线工艺点位约 3 万点，数据具有变化快、通讯量大、存储周期长等特点，采集范围涵盖以上三大类系统。通过 OPC DA、Modbus TCP、IEC 104 等工业协议实现与国内外主流 DCS、PLC、电力综保的实时采集，数据采集与处理过程需确保及时性、完整性、可靠性，一般数据更新周期小于 100ms，数据吞吐量不小于 2 万点/秒。为保证控制系统的安全隔离，可以从

工厂中的实时数据库系统中导出历史 3-5 年的工艺运行数据，结合工艺专家与操作工的经验，确定研究对象中涉及装置、核心工序与关键设备，遴选温度、流量、压力、电流、电压等重要工况参数，同时对 D 罐中的胶粒浓度化验结果进行搜集，并与工艺数据导出的时间进行对齐，最终数据样本如下所示。

表 6 样本数据表

	date	ZI_A50701	FI_A50202	TI_50211	.....	SC_AL502	CI_A50261	label
1	2017-10-03 00:00	48.725	3912.4458	63.377823	.....	100	True	16.01
2	2018-10-03 00:01	48.7261	3946.283	62.902885	.....	100	True	16.01
3	2018-10-03 00:02	48.8355	3935.5342	62.59909	.....	100	True	16.01
4	2018-10-031 00:03	48.7237	3975.6162		.....	100	True	16.01
5	201810-03 00:04	48.852	3848.2122	62.98378	.....	100	True	16.01
6	2018-10-03 00:05	48.7428	3901.7546	61.2832	.....	100	True	16.01
7	2018-10-03 01:55	48.7223	3950.3738	59.2169	.....	100	True	16.18
8	2018-10-03 01:56	48.7613	3793.2427	58.7469	.....	100	True	16.18
9	2018-10-03 01:57	48.7227	3946.152	60.8946	.....	100	True	16.18
10	2018-10-03 01:58	48.7633	3956.1472	62.1963	.....	100	True	16.18
11	2018-10-03 01:59	48.7861	3939.103	63.8705	.....	100	True	16.18
12	2018-10-031 02:00	48.7933	3677.3767	62.175	.....	100	True	16.18
13	2018-10-03 02:01	48.7543	3954.5867	61.957	.....	100	True	16.18
14	2018-10-03 02:02	48.8204	3905.8347	62.2076	.....	100	True	16.18
15	2018-10-03 02:03	48.7334	3913.8547	62.5133	.....	100	True	16.18

	date	ZI_A50701	FI_A50202	TI_50211	.....	SC_AL502	CI_A50261	label
16	2018-10-03 02:04	48.716	3915.5413	63.8572	.....	100	True	16.18
17	2018-10-03 02:05	48.8165	3961.5615	61.6341	.....	100	True	16.18
				.....				
1554121	2020-09-30 00:00	51.7758	3906.7046	59.4615	.....	100	True	15.83

3、数据准备

（1）滤波降噪

实际生产数据波动较大，伴随周期性噪声，需要对这些干扰进行抑制。

引入滑动平均滤波方法，设置长度为 N 的窗口滤波器，把连续取 N 个采样值看成一个队列，采用先进先出原则，每次采样到一个新数据放入队尾,并扔掉原来队首的一次数据。对窗口中的数据进行算术平均值计算，获得滤波后的数据。从而避免噪声数据对后续模型训练的干扰。如图所示，II_A_L502 电流位号的前后滤波示意图如下。

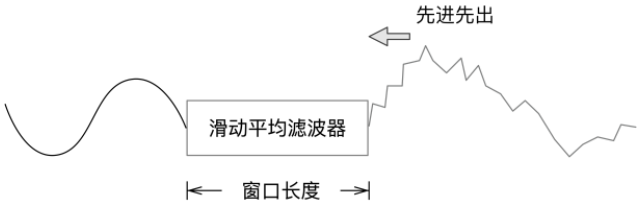


图 40 滑动平均滤波器原理图

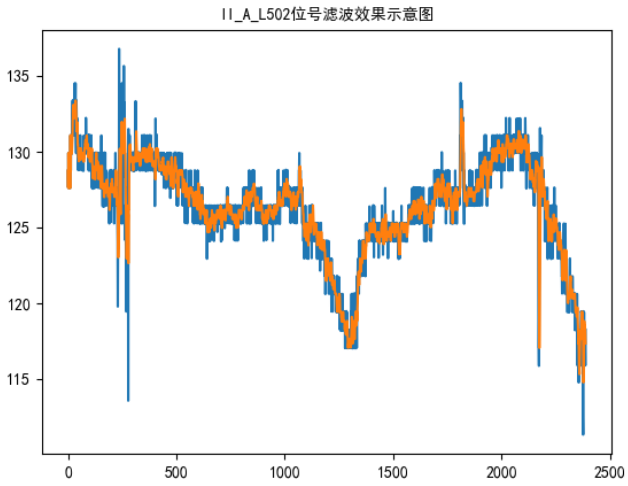


图 41 II_A_L502 位号滤波示意图

（2）缺失值处理

DCS 采集控制系统有上万数据点同时接入，大量数据写入 PI 数据库中，尽管京博橡胶采用高稳定性的 DCS 采集控制系统和 PI 实时数据库，但也会因为以下原因导致样本数据中存在数据缺失:

- 异常工况导致数据收集失败，比如传感器、DCS 采集器异常
- 机械原因所导致数据保存失败，比如数据存储的失败，存储器损坏，机械故障导致某段时间数据未能收集。

如果样本数据中有值为空值或者 NaN,将会导致算法无法处理这些值，造成算法程序执行异常。

因此，需要对样本集中的数据缺失进行处理，本课题样本中缺失数据绝大多数呈离散型分布，极少数位号点连续缺失。为此采取的策略对离散型分布缺失值进行临近值填充; 对极少数时间连续缺失的数据,进行整行删除。这样最大化地避免了数据浪费，

又有效地对缺失值进行了填补。

图 42 个别缺失与大面积缺失数据展示

（3）归一化

样本数据中会发现不同特征之间的数值量程相差较大，比如流量、压力的量程远远大于电流、温度等位号的量程，这样一些数值大的数据，一方面会造成计算数值较大，速度慢；另一方面，较大的量程差异会导致在后续模型训练中不容易收敛，模型训练时间长。

因此，采取归一化处理，具体策略为在已知各个位号传感器量程的情况下，以量程为区间范围，把数据映射到 0～1 范围之内。

$$x' = \frac{x - min}{max - min}$$

其中：

min 为量程最小值；

max 为量程最大值；

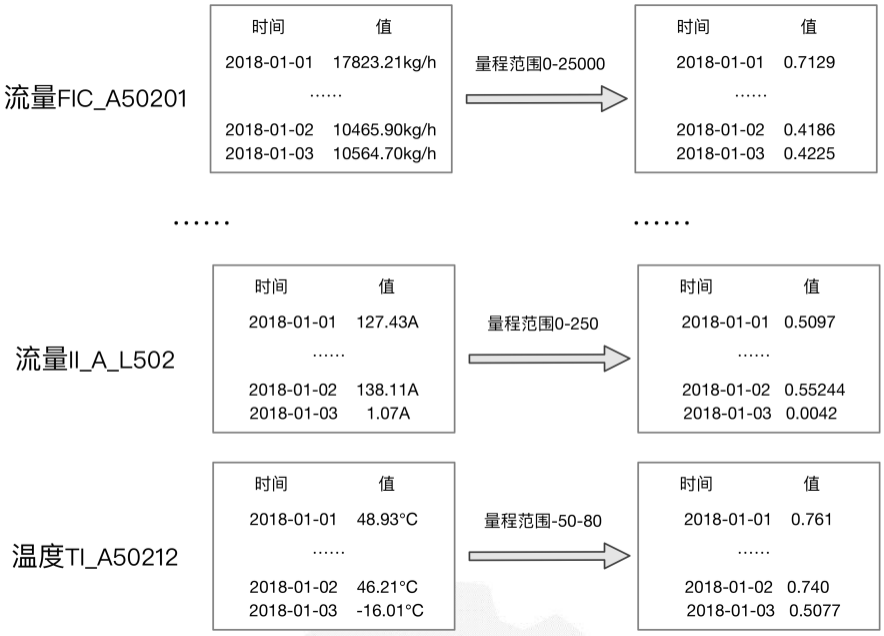


图 43 不同数据的归一化处理

（4）相关性分析

特征选择作为数据科学中及其重要的一环，如果相关分析时各自变量跟因变量之间没有相关性，就没有必要再做回归分析；如果有一定的相关性了，然后再通过回归分析进一步验证他们之间的准确关系。

通过特征选择的降维方法降低了建模过程中的计算复杂度、并排除干扰特征项提高模型精度。

本案例采用 Pearson 相关性系数（Pearson Correlation）对上述 40 余特征进行相关性分析。

Pearson 相关性系数（Pearson Correlation）是衡量向量相似度的一种方法，输出范围为-1 到+1，0 代表无相关性，负值为负相关性，正值为正相关，公式如下：

$$\rho(X,Y)=\frac{\mathrm{E}[(X-\mu_x)(Y-\mu_Y)]}{\sigma_X\sigma_Y}=\frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i-\mu_Y)^2}}$$

分析结果梯形图如下所示：

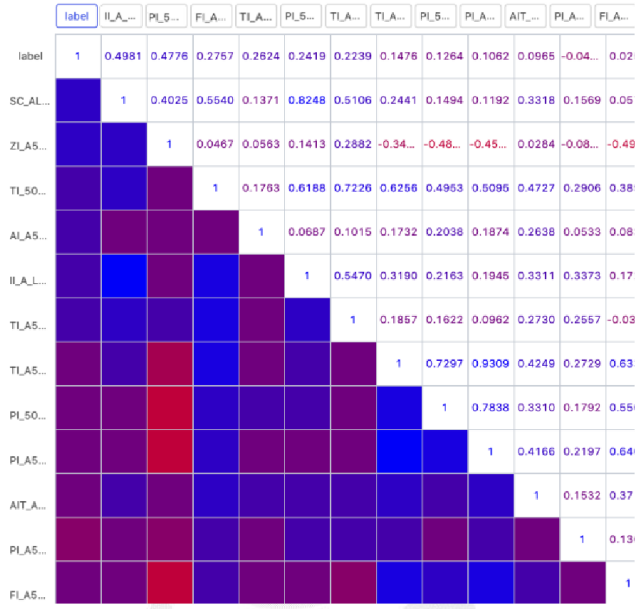


图 44 相关性分析结果示意图

根据相关性分析可知，下料流量 FIC_A50103，挤压机的电流 II_A_L502，电机温度 TI_A50212，磨头压力 PI_A50208，502 罐进料压力 PI_A50202，502 罐温度 Ti_A50202 等 6 个参数相关系数大于 0.2。

结合相应的人工经验得知，橡胶原液通过下料口进入挤压机前，一般其水质量分数约为 95%，进过脱水挤压机和将水质量分数降至 50%，再通过后续的已烷注入混合成橡胶浆液。

因此，当下料口的橡胶粒变化的同时，挤压机的相关工艺参数也会随之变化，比如下料量中橡胶较多，伴随挤水量也增多导致挤压机的功率因此上升；下料量中橡胶较少，挤压机的功率也

下降。

因此，特征相关性分析和工艺经验，将特征重点聚焦在挤压机的工作参数范围内，有温度下料流量 FIC_A50103，挤压机的电流 II_A_L502，电机温度 TI_A50212，磨头压力 PI_A50208 共计四个。

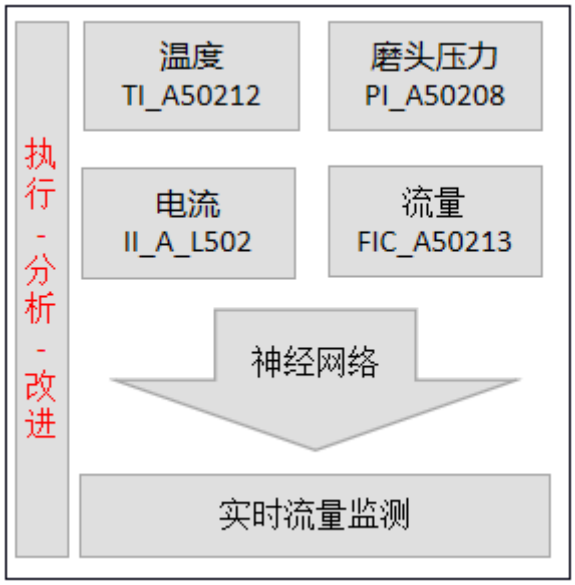


图 45 特征参数数据

4、数据建模

（1）建立神经网络预测模型

建立 DNN 深度神经网络：神经网络为 4 层，输入层神经元节点为 4，激活函数为 relu；隐藏层 1 神经元为 20，激活函数 relu，设置 dropout 系数，随机丢弃网络中的神经元；隐藏层 2 神经元为 16，激活函数 relu；输出层神经元为 1。从而建立设备参数与挤压机脱水率的映射模型，取得初步良好效果，模型框图如下图所示。

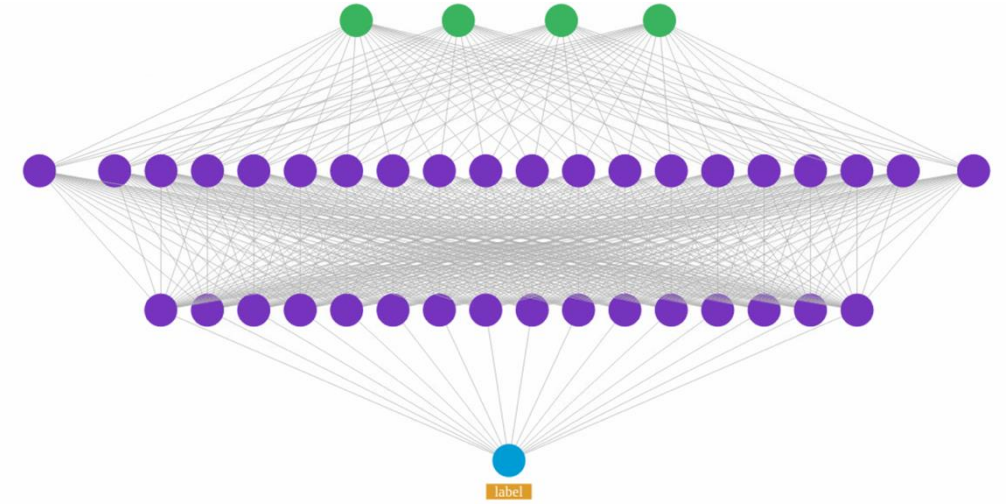


图 46 神经网络预测模型示意图

（2）建立机器视觉模型

机器视觉是通过机器视觉产品（即图像摄取装置，分 CMOS 和 CCD 两种）将被摄取目标转换成图像信号，传送给专用的图像处理系统，根据像素分布和亮度、颜色等信息，转变成数字化信息，图像系统对这些信号进行各种运算来抽取目标的特征，进而根据判别的结果来控制现场的设备动作。

它是计算机学科的一个重要分支，它综合了光学、机械、电子、计算机软硬件等方面的技术、设计到计算机、图像处理、模式识别、人工智能、信号处理、光机电一体化等多个领域。

我们通过 BP 神经网络建立了一个生产过程中工艺参数、预测值的映射模型，准确的说是挤压机相关工艺参数与橡胶颗粒量的映射模型，但在生产环境变化的情况下，预测值波动明显，且存在一定误差，因此本案例设计在橡胶原液增加或下降过程中，使用机器视觉检测方案来辅助，具体研究内容如下，在上述条件

准备完成后，得到如下的图像:

1）区域定位

显而易见，橡胶原液直接通过左侧振动带直接进入下方下料口中，对我们来说，真正需要的监测的地方只有左侧振动带区域，其他的地方反而会影响预测精度，因此首先需要对指定监测区域进行定位截取，挑选出如下区域:

图 47 监测区域定位截取

这样就涵盖了绝大部分的下料量画面，从而也避免了其他噪声的干扰。

2）掩膜

图 48 形成灰度图

3）特征提取

得到上述处理过后的灰度图后，统计其灰度分布曲线图、灰度分布直方图，如下图所示，可以较为清晰地区分出了振动传送带上橡胶粒与其他不同区域的灰度值，其中振动传送带上橡胶粒集中分布在灰度值155-205之间，因此通过截取该范围的灰度值，处理后结果如下

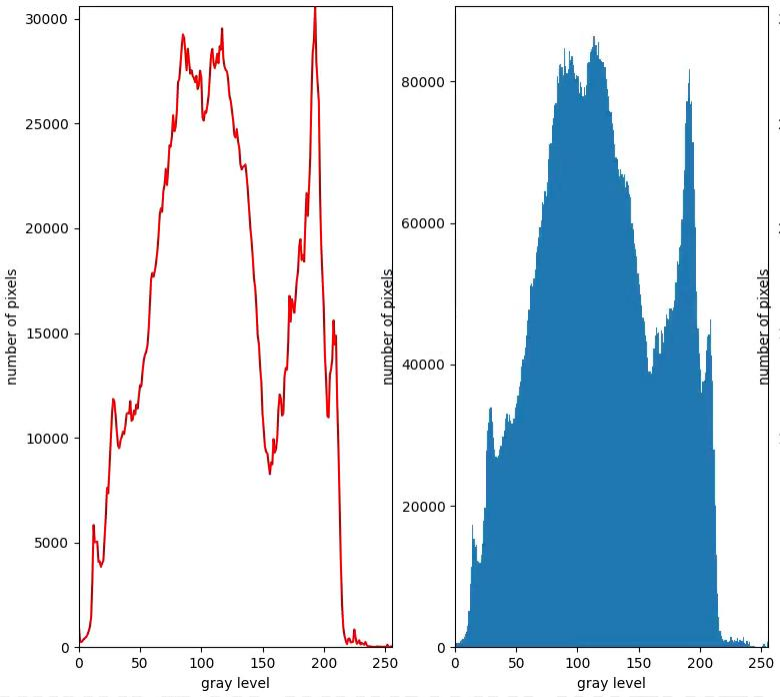


图 49 灰度分布统计

图 50 范围截取

4）数值计算

再计算指定区域内白色像素点的大小，作为视觉算法的输出。

5、模型验证

当橡胶颗粒下料量稳定的情况下，数据挖掘模型效果优于视觉模型，当下料量上升或下降，处于不稳定的情况下，视觉模型效果优于数据挖掘模型，因此非常符合集成学习的思想，通过判断下料量的稳定与否，分别采取不同的模型来预测，从而提高整体系统的稳定性。

剩下的就是如何判断下料量是否稳定，需要建立一个相应的判断指标，实际生产中发现，当下料量开始稳定的时候，振动传送带表面橡胶粒逐渐饱和，开始厚度堆积，而不稳定的时候振动传送带表面或多或少会出现空缺面积，另外，由于传送带的振动特性，橡胶胶粒被均匀的铺在传送带上，肉眼可以直接观测出实际流量变化，因此可以根据橡胶胶粒铺满整个振动传送带为临界点，来选择切换上述两个不同模型。

$$\begin{cases}\text{采用数据挖掘模型，工况稳定时}\\ \text{采用机器视觉模型，工况不稳定时}\end{cases}$$

其中将 5600kg 作为工况稳定的分界线的话，实际模型切换模式如下:

$$\begin{cases}\text{采用数据挖掘模型，机器视觉检测值} > 5600\text{kg/h}\\ \text{采用机器视觉模型，机器视觉检测值} \leq 5600\text{kg/h}\end{cases}$$

通过该方案上线运行，能够 24 小时在不同给的情况下对下料口的实时干胶量进行预测，指导现场生产。

（三）实施效果

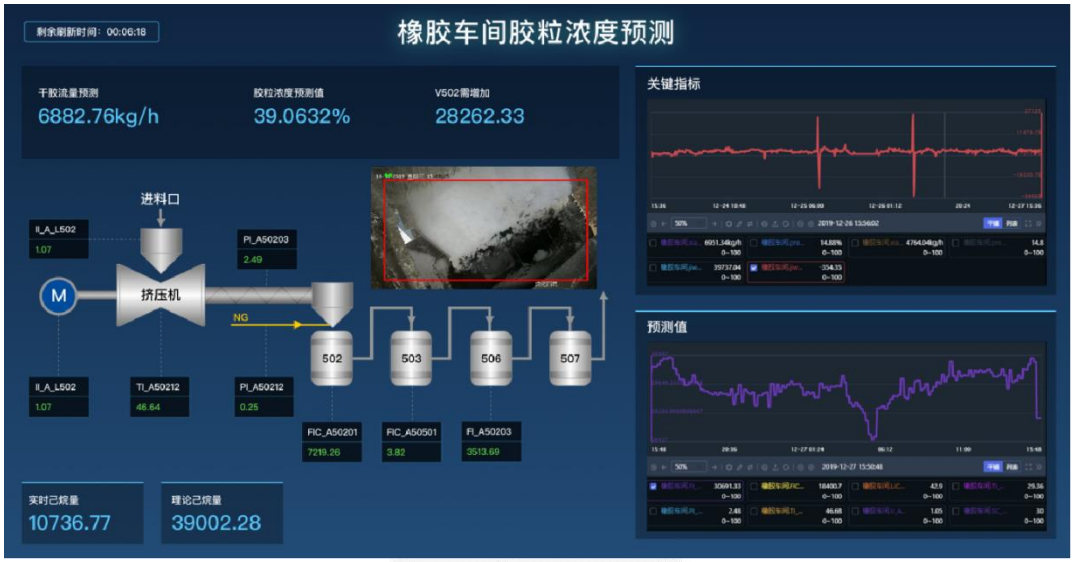


图 51 系统截取

（1）精度分析

选取京博橡胶 5 万吨/年丁基橡胶装置为应用实施对象，结合现场操作工经验，建立 BP 神经网络回归预测模型，输出稳态工况下的胶粒浓度、已烷添加量预测值；利用机器视觉方法辅助来料工况判定，提升回归预测模型的鲁棒性，从而为操作工提供 24 小时实时在线配比建议。

橡胶浓度场景经过一年的优化，基本消除以往化验室反馈调整的工艺滞后性，并且由原先的单一人工经验转变为人工+预测模型的生产指导方式。实现在不同工况下干胶流量、胶粒浓度的预测达到预期精度（干胶流量误差在 50 公斤内，浓度误差在 5‰），进一步给出 502、503、506 罐的正己烷配比量推荐。橡胶浓度 CPK 由原来的 0.95，上升至 1.02，最终产品质量稳定性提升了 2.4%，带来上百万的经济效益，极大地提高了产品竞争

力。

（2）适应性分析

橡胶生产工艺相对成熟、趋于稳定，对于不同产线的复制与推广来说，更多的变数存在于核心设备机械性能差异、原材料质量以及整条产线的效能不同产线由于设备性能退化、工艺调整，从而导致原有模型精度不再满足要求。通过 supOS 工业操作系统对关键特征的长期监测采集，实现大数据和机器视觉模型的自学习，更加有效地促进回归预测模型进行持续优化，从而适应不同的生产工况变化，建立中长期稳定模型，定期替换原有模型从而保证原有模型精度。形成可复制、可推广的工业智能场景应用创新方法，也可以面向不同的卤化丁基橡胶工厂以及不同行业、不同领域进行推广。

五、结语与展望

工业大数据分析与应用中存在着场景多样、数据量大、业务逻辑复杂、影响因素繁多、耦合关系紧密、技术难度大等诸多难题，虽然针对不同的业务需求及应用场景往往需要“一事一议”，但是看似在不同行业企业中、面向不同业务域、采用了不同技术手段的各个成功应用实践之中仍然是有相通之处的，特别是其采用的方法论。

我们曾在《工业大数据分析指南》中详尽的介绍过这种工业大数据分析与应用方法论，其分析过程包括业务理解、数据理解、数据准备、建模、验证与评估、实施与运行等六个基本步骤。通过对本白皮书遴选四个典型案例的深度剖析可见，各个案例都不约而同地遵循了该方法论，无一不是基于深刻的业务理解，进而到数据理解，对数据进行各种预处理，再通过对于数据的组织和建模凝练和固化知识，最终形成以为模型为核心的应用。这也反复证明了该方法论对具体应用是有较强指导意义的，值得在实践中加以推广。

当然，我们编写本白皮书并不是意在验证之前提出的方法论，而更多的是向广大技术研究及工程实践人员介绍在工业大数据分析与应用领域内的优秀做法，所选取的四个典型案例也是从众多应用案例中经过了仔细遴选，每个案例都具有很强的典型性和各自特色，通过对其成功经验分享，既可开阔思路，形成启发，亦可促进推广应用。

当前，随着大数据、云计算、物联网、人工智能等为代表的

新一代信息技术的不断成熟，工业大数据作为制造业提升生产力、竞争力和创新力的关键因素，正在推进制造业发展向智能化新模式的转变。在工业大数据技术未来的应用落地中，有两个问题尤其值得关注。

首先，我们必须要承认工业大数据整体的技术还在高速发展中，数据体系建设不完善、数据驱动方法的因果性缺失等问题使得分析的准确性、结果的泛化性、应用的确定性可能都无法达到一个非常理想的水平，那么我们应该如何在有限的技术水平下更好的结合业务场景落地？这需要技术和业务两端发力，一方面技术不断发展并向工业场景渗透应用，另一方面，业务问题也需要不断拆解、细化，业务上的使用场景要不断地探索与磨合，作为使用者的业务人员也需要不断的与算法磨合，了解算法的应用优势与缺陷。在人与算法、业务与技术的不断应用、适配与迭代发展中，工业大数据分析才能得到长足的进步与持续的发展。

第二个问题是数据驱动的分析方法与机理驱动的正向设计的深度融合。伴随着我国制造业企业不断升级及基础研究取得突破，我们对于各事物的了解不断深入，特别是对于物理世界的运行机理将会有更加深刻的认识，掌握了生产制造的客观规律，将有助于解决当前生产中仍存在的一些问题。随着工业机理认知的深入和技术手段的不断发展，工业大数据分析不再仅关注于数据本身的建模，而是在挖掘数据特征背后的物理意义以及特征之间关联性机理上会扮演越来越重的角色。我们可以这样总结，机理是基础，数据是资源，而工业大数据技术正是不断通过对资源的

挖掘来持续夯实基础的手段。

可以预见到，在不久的将来，数据价值将会充分放大，数据将成为企业宝贵的资源之一，各类真正具备智能化特征的应用将会不断涌现，工业大数据必将进入高速发展的快车道。