大数据驱动的质量管理：基于机器学习的产品缺陷预测

段景阔[1] 宋建于（通讯作者）[2]

1. 四川鸿歌网络技术有限公司 四川省南充市 637500

2. 蜂巢能源科技股份有限公司 江苏省常州市 213212

【摘　要】：针对离散制造过程中质量波动诱因复杂、缺陷追溯滞后的问题，本文以电子制造业 SMT 生产线为研究对象，构建基于 XGBoost 的产品缺陷预测模型。研究整合贴装精度、吸嘴负压、回流焊温度曲线、焊膏检测结果及 AOI 判定记录等多源数据，通过样本对齐、缺失值补全、异常值识别和互信息特征筛选形成结构化输入，并采用网格搜索与 5 折交叉验证优化模型参数。测试结果表明，优化后模型准确率达到 92.5%，召回率达到 89.3%，F1 值达到 90.8%，较优化前准确率提高 4.4 个百分点。实际应用显示，该模型可支持缺陷预警与工艺调整，使 SMT 焊接缺陷率降低 15%，生产成本节约 8%，具有较好的质量预防与现场应用价值。

【关键词】：大数据；质量管理；缺陷预测；XGBoost；SMT 生产线

0 引言

随着离散制造向柔性化、定制化和智能化方向发展，传统依赖事后检验的质量控制方式难以及时应对多工序耦合、设备状态波动和物料批次差异带来的质量风险。SMT 生产线数据采集密集、缺陷类型多样，具备开展数据驱动质量预测的现实基础。研究机器学习在焊接缺陷预测中的应用，有助于提升质量管理的前置性、精准性和可追溯性。

1 大数据驱动的质量管理与产品缺陷预测概述

鉴于离散制造过程呈现多工序耦合、设备状态波动以及质量追溯链条延长等特性，质量管理正在由事后检验逐步转向数据驱动的过程预防。大数据驱动的质量管理，是指把生产设备传感数据、工艺参数、检验记录、物料批次以及售后反馈等多源信息纳入统一分析框架，借助数据清洗、特征提取以及关联建模来识别质量波动的潜在诱因。产品缺陷预测则是在上述数据基础上，对可能出现的尺寸偏差、性能失效以及外观异常开展概率化判别，为工艺调整、设备维护以及供应链协同提供前置依据。从制造现场运行逻辑来看，该方法并非替代传统质量控制，而是把经验判断转化为可计算、可追溯的决策支持，使质量管理活动更契合柔性生产以及批量定制场景下的快速响应需求。

2 基于机器学习的产品缺陷预测模型构建

2.1 缺陷预测数据预处理方法

电子制造业 SMT 生产线的焊接缺陷预测数据具有高频采集、多设备关联以及批次差异明显等特性，数据预处理需要围绕焊点形成机理来开展。本文把贴片机视觉系统记录的贴装精度、吸嘴负压、元器件偏移量，回流焊炉采集的升温斜率、峰值温度、保温时间，以及元器件来料参数、焊膏批次、SPI 锡膏检测结果和 AOI 外观判定记录纳入同一数据表征空间。鉴于设备日志与检验记录存在时间粒度不一致的问题，可依靠工单编号、PCB 条码以及炉温曲线时间戳进行样本对齐，把单个焊点或单块 PCB 作为建模样本。缺失值处理需区分形成缘由，传感器瞬时丢包可选用相邻时间窗口均值进行填充，批次字段缺失则运用同工单物料记录进行补全。异常值处理不宜简单删除，可结合 3σ 准则以及箱线图识别超出工艺窗口的温度、压力和偏移数据，并且把真实工艺漂移保留下来。特征工程环节运用互信息法衡量工艺变量与桥连、虚焊、立碑等缺陷标签之间的依赖强度，筛选回流焊峰值温度、保温时间、贴装 X 向偏移量、焊膏厚度均值等关键特征，从而把原始制造数据转化为契合模型训练需求的结构化输入。

2.2 机器学习预测模型的选择与训练

SMT 焊接缺陷样本往往呈现非线性边界以及类别不均衡特性，模型选择需要兼顾可解释性、拟合能力以及现场部署成本。逻辑回归结构清晰，能够反映单一工艺变量对缺陷概率的方向性影响，但面对回流焊温度曲线、贴装偏移以及元器件尺寸之间的交互关系时，线性假设会压缩模型表达空间。随机森林依靠多棵决策树进行集成判别，对噪声字段具有较好鲁棒性，不过树结构数量增加后，缺陷边界附近样本的概率校准能力会受到限制。XGBoost 把梯度提升树当作核心学习器来使用，可在逐轮迭代中修正前一轮残差，并且借助正则项约束树复杂度，较适宜处理 SMT 生产线中多工艺参数耦合的缺陷预测任务。结合训练流程，原始样本经清洗、特征筛选以及标签编码后，按照 7:3 划分训练集和测试集，在训练阶段设置学习率为 0.1、最大树深度为 5、子采样比例为 0.8，并把二分类逻辑损失作为优化对象。模型流程如图 1 所示，其核心链路体现了从多源

制造数据输入到缺陷概率输出的完整建模过程。XGBoost 训练过程可用目标函数表示如下：

$$\mathcal{L}^{(t)}=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\sum_{k=1}^{t}\Omega(f_k),\Omega(f)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$$

其中，$\mathcal{L}^{(t)}$ 表示第 t 轮迭代的目标函数，n 表示训练样本数量，l 表示缺陷标签与预测结果之间的损失函数，y_i 表示第 i 个样本的真实缺陷标签，$\hat{y}_i^{(t-1)}$ 表示前 $t-1$ 轮模型输出，$f_t(x_i)$ 表示第 t 棵树对样本 x_i 的预测增量，$\Omega(f_k)$ 表示树模型复杂度惩罚项，γ 表示叶节点数量惩罚系数，T 表示叶节点数量，λ 表示叶节点权重正则化系数，w_j 表示第 j 个叶节点权重。

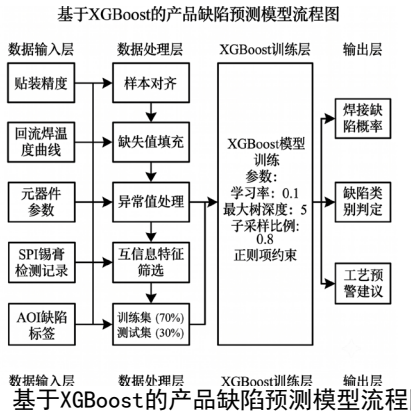

图1 基于XGBoost的产品缺陷预测模型流程图

2.3 模型优化策略

模型初始训练完成后，参数组合对 SMT 焊接缺陷识别边界的影响会逐步显现，单一经验设定难以覆盖不同产品批次、焊盘尺寸以及炉温窗口带来的样本差异。围绕 XGBoost 模型优化，可把学习率、最大树深度、最小叶节点样本权重、子采样比例以及树列采样比例设置为候选参数空间，运用网格搜索法进行组合遍历。该方法把每一组超参数都置于相同训练数据条件下进行比较，能够减少人工调参带来的主观偏差。考虑到 SMT 缺陷样本中正常焊点数量通常高于缺陷焊点数量，调参过程还需关注类别权重以及评价指标稳定性，避免模型过度偏向多数类别。为检验模型在不同样本切分条件下的泛化能力，可把训练集划分为 5 个互斥子集，轮流选取其中 1 个子集作为验证集，其余 4 个子集参与模型拟合，并且把 5 次验证结果取均值作为参数组合筛选依据。该优化思路并非单纯追求训练集拟合程度，而是把跨批次波动、设备状态微变以及来料差异纳入模型稳健性考量。5 折交叉验证准确率可表示为：

$$\overline{Acc}_{cv}=\frac{1}{K}\sum_{k=1}^{K}\frac{TP_k+TN_k}{TP_k+TN_k+FP_k+FN_k}$$

其中，$\overline{Acc}_{cv}$ 表示交叉验证平均准确率，K 表示交叉验证折数，在本研究中取值为 5，TP_k 表示第 k 折中被正确判定为缺陷样本的数量，TN_k 表示第 k 折中被正确判定为正常样本的数量，FP_k 表示第 k 折中被误判为缺陷样本的正常样本数量，FN_k 表示第 k 折中被误判为正常样本的缺陷样本数量。

3 缺陷预测模型的效果评估与应用实践

3.1 模型效果评估指标与结果分析

SMT 生产线缺陷预测模型的评价不能仅停留在总体判别正确率层面，还需要把缺陷样本被识别出来的比例以及误判代价纳入同一分析框架。鉴于桥连、虚焊以及立碑等缺陷在样本分布中占比相对较低，本文选用准确率、召回率以及 F1 值开展综合评价，其中准确率反映模型对正常焊点以及缺陷焊点的整体区分能力，召回率用于衡量缺陷焊点被捕获的充分程度，F1 值则把查全能力以及判别稳定性进行折中表达。在测试集评估中，优化后 XGBoost 模型准确率达到 92.5%，召回率达到 89.3%，F1 值达到 90.8%，相较优化前 88.1% 的准确率提高 4.4 个百分点。结合表 1 可以看出，参数优化后模型在漏判控制方面表现更为稳健，说明网格搜索以及 5 折交叉验证能够缓解单一参数设定对缺陷边界识别带来的偏差。

表1 SMT生产线缺陷预测模型性能对比表

模型阶段	准确率/%	召回率/%	F1值/%	误报率/%	漏报率/%	性能说明
优化前XGBoost模型	88.1	84.7	86.2	9.6	15.3	对多数正常样本识别较稳定，但对低占比缺陷焊点存在漏判
优化后XGBoost模型	92.5	89.3	90.8	6.8	10.7	缺陷捕获能力以及综合判别能力同步提高，误报与漏报得到压缩

3.2 模型在实际生产中的应用价值

在 SMT 生产线连续运行场景下，缺陷预测模型的价值主要体现在把检验节点后移造成的质量损失转化为工艺过程中的前置干预。模型接入贴片机、SPI 检测设备、回流焊炉以及 AOI 检测系统后，可对贴装偏移量、焊膏厚度均值、峰值温度以及保温时间等变量进行在线判别，并且把缺陷概率高于预警阈值的 PCB 批次推送至工艺工程师工作台。结合实时预警信息，现场可围绕吸嘴清洁、钢网状态、炉温曲线以及来料批次开展针对性调整，减少桥连、虚焊以及立碑缺陷在后续批次中的延续扩散。从经济效益角度来看，模型运行后 SMT 焊接缺陷率降低 15%，生产成本节约 8%，其收益主要来自返修工时压缩、报废板减少以及停线排查时间缩短。需关注的是，模型对新产品首

批数据依赖较强，后续可引入在线增量学习以及漂移监测机制，使模型更好适配多品种小批量生产节奏。

4 结语

本文围绕 SMT 焊接缺陷预测构建了大数据驱动的质量管理路径，验证了 XGBoost 模型在多源工艺数据融合、非线性缺陷边界识别和现场预警中的适用性。通过参数优化与交叉验证，模型在准确率、召回率和 F1 值方面均取得提升，可为工艺调整、设备维护和批次管控提供依据。后续研究可进一步结合在线增量学习、数据漂移监测和多产品迁移建模，增强模型在小批量、多品种生产环境中的持续适应能力。

参考文献

[1]邱云峰,李泽宏.基于超参数优化机器学习算法与BP神经网络模型的元器件质量监测与故障预测研究[J].电子元件与材料,2025,44(10):1237-1244.

[2]喻皓,张莹,李倩,姜立标,尚云鹏.基于机器学习的软件缺陷预测研究[J].重庆大学学报,2025,48(02):10-21.

[3]刘大千,王共冬,许成阳,毕晓萱.基于经验知识的机器学习CFRP/AL叠层材料制孔质量分层预测[J].科技和产业,2025,25(13):22-27.

[4]程伟,赵永强,庞嘉尧,何勇,余乐.基于机器学习的熔融沉积成型表面质量预测控制研究进展[J].中国塑料,2025,39(08):131-138.

[5]邱勇.基于集成学习的TFT-LCD质量预测方法研究[D].导师：任鸿儒;杨梅.广东工业大学,2023.

作者简介：段景阔（1971—），男，汉族，大专，研究方向为大数据。