基于 CART 算法的管线钢铸坯探伤结果预测与工艺诊断

王复越[1,2]，任　毅[1,2]，田永久[3]，崔福祥[3]，张哲睿[1,2]，付成哲[1,2]

（1.海洋装备用金属材料及其应用国家重点实验室，辽宁 鞍山 114009；2.鞍钢集团钢铁研究院，辽宁 鞍山 114009；3.鞍钢股份有限公司鲅鱼圈分公司，辽宁 营口 115007）

摘　要：本研究针对超洁净管线钢产品需求，采集某中厚板生产线管线钢连铸坯生产数据，结合冶金学原理和皮尔逊相关系数选取关键工艺特征属性，以管线钢板的探伤结果为目标标签，采用决策树算法建立管线钢连铸坯质量预测模型。经过对模型结构的调整和优化模型评价，得到了预测效果好（测试集的 AUC 值为 0.848）、泛化能力强（△AUC 值为 0.042）的模型。本决策树预测模型提供了一种简单、高效、可解释性较强的预测方法。此外，根据重要度得分准确定位了工艺的关键点以及阈值。该方法的应用为企业工艺智能调整和产品质量智能管理提供帮助。

关键词：决策树；机器学习；管线钢；连铸；预测模型

Abstract: This study aimed at the need for ultra-clean pipeline steel products. The production data of pipeline steel billet was collected and used in an iron and steel enterprise, combined with the principle of metallurgy and Pearson correlation coefficient selection of crucial process characteristics of casting properties, for inspection results of pipeline steel plate as the label, the decision tree algorithm is adopted to establish the quality prediction model of pipeline steel continuous casting billet. After adjusting the model's structural parameters and model evaluation, the model with good prediction effect (AUC value of the test set is 0.848) and good generalization ability (△AUC value is 0.042) was obtained. The decision tree prediction model generated in this study provides a simple, efficient, and intense interpretation of the results of the prediction method. The key points and thresholds of the process were accurately located according to the importance score. It provides help for intelligent adjustment of enterprise processes and intelligent management of product quality.

Key words: Decision Tree Algorithm; Machine Learning; Pipeline Steel; Continuous Casting; Prediction Model

0　引言

近年来，国家倡导要充分发挥海量数据和丰富应用场景优势，促进数字技术与实体经济深度融合，赋能传统产业转型升级，不断做优、做大我国数字经济[1]。新一代先进钢铁材料的研发与钢铁行业的发展应在现有知识与理论框架下，利用钢铁行业自动化程度高、工业数据完整性好的特点，充分发挥集成计算材料工程及材料信息学的优势，创建智慧研发设计路线，结合现有数据开发高效预测与评价办法，进而实现中国先进钢铁材料的研发与产生基于智能创新的发展[2]。目前，认为人工智能是钢铁行业进行大数据分析及数据挖掘工作的必然选择，机器学习是实现人工智能的一种重要方式。而机器学习领域已经发展出诸多适用于不同场景的计算方法，如决策树、支持向量机、贝叶斯学习、随机森林、人工神经网络等[3,4]。经过这些年的发展，机器学习算法建模已与钢铁行业各环节的实际生产有着广泛的结合，可以应用在各项产品性能指标的预测，辅助钢铁生产过程的决策，并与传统的科研研发、生产管理路径相辅相成[5-7]。

高级别管线钢为满足众多服役性能一般要求管线钢铸坯应具有高纯净度、精细组织和严格的夹杂物控制[8]。然而管线钢铸坯的生产要经历转炉炼钢(BOF)、钢包精炼(LF)、真空脱气(RH)、钙处理、连铸、末端轻压下等多个工艺环节，流程长、各关节控制关键点多，不同环节间相互制约、相互影响。目前，国内外对管线钢铸坯质量的判定处在人工判定阶，而人工判定存在效率低、探伤抽检数量大、漏检率高、判定结

果滞后等弊端。

为解决上述问题，本文提出一种基于决策树算法思想，使用 Python 编程语言在 Pycharm 集成开发环境下建立并运行的管线钢连铸坯质量预测模型，通过优化模型参数实现管线钢铸坯质量的智能判定，提高质量判定准确率，提升出厂产品合格率。此方法有助于实现炼钢及铸造环节工艺的优化，对于提升设备运行水平以及提高产品质量也具有重要的意义。

1 建模方法

1.1 数据集处理与特征属性筛选

本研究采集了 2897 组管线钢铸坯的工业数据，每组数据代表一个管线钢铸坯，由铸坯的化学成分、RH 处理时间、浇铸调度、浇铸速度、浇铸温度、轻压下辊间隙等 15 个生产工艺参数（特征属性）组成。在进行特征属性筛选之前，先对每个特征属性的范围映射为[0,1]，以减少特征属性间量级的差异。各备选特征的均值、最小值、最大值和归一化方差如表 1 所示。剔除空白值和乱码数据后，按照 8∶2 的比例将建模数据集随机划分为互斥的训练集与和测试集。

表 1 数据集信息

Table 1 Information for the datasets

序号	特征	简写	最大值	最小值	均匀值	归一化方差
1	T.S, ppm	S	30	7	23	0.008
2	T.Al, ppm	Al	747	321	479	0.026
3	T.Ca, ppm	Ca	23	6	8	0.005
4	T.O, ppm	O	523	12	16.2	0.012
5	T.N, ppm	N	53	32	32.5	0.007
6	RH 处理周期, min	Rp	36	25	26	0.002
7	RH 净处理时间, min	Rnt	26	10	14.6	0.043
8	氩气流量, ml/s	Afr	88.0	19.9	52.5	0.016
9	静置时间, min	St	9.7	2.2	5.4	0.003
10	铸造过热度, C°	Coh	47	15	32	0.028
11	二冷水流量, $L/m^2 \cdot s$	Con	3.4	3.3	3.32	0.003
12	铸造速度, mm/min	Cs	1.04	0.65	0.86	0.048
13	冷却水温度, C°	Cwt	20.4	12.3	17.3	0.026
14	水口服役时长, h	Stn	1	9	4.6	0.001
15	轻压下辊缝合格率, %	Qrg	95.73	90.23	93.89	0.016
16	缺陷数量	Nd	5	0	0.056	0.062

为避免了“维数灾难”，本研究采用数据驱动的降维方法。从表中可以看出，RH 处理周期、氩气流量、静置时间、二冷水流量、冷却水温度、喷嘴使用时间以及化学成分中的 S、Ca、N 等 9 个熔炼和铸造参数的归一化方差均小于 0.01，这说明这些参数在数据分布相对集中，对铸坯冶金质量的影响相对较小，因此可以剔除这些特征。进一步使用皮尔逊相关性来描述特征之间以及特征与缺陷处置数量之间的相关性，如图 1 所示。可以看出，T. O、水口服役时长和缺陷处理数量之间的相关性均小于 0.1，表明这些特征与目标标签的相关性较小。最后，确定管线钢连铸坯探伤模型的特征设置如为：RH 处理周期、钢中总铝含量、RH 净处理时间、钢水浇铸过热、浇铸速度、轻压下辊缝间隙合格率六项。

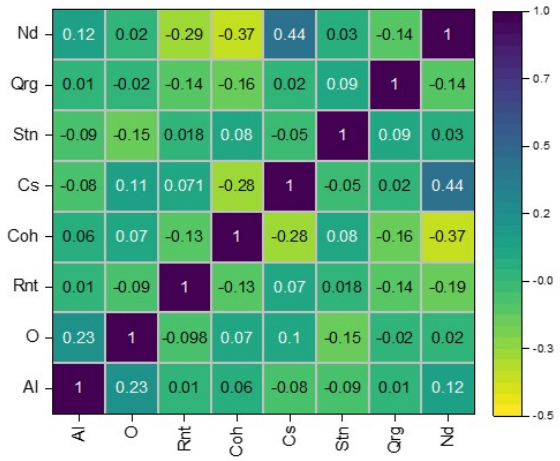


图 1　特征属性及缺陷数量间皮尔逊相关性热力矩阵

Fig. 1　Pearson's correlation between features and target label

1.2　模型建立与评估方法

构建决策树的主要思想是选用不同的样本纯度度量指标（信息增益、增益率、基尼指数）找到包含关于目标特征的最大信息量（纯度）的描述性特征[9]。然后沿着这些特征的值分割数据集，使得生成的子数据集中的目标特征值纯度尽可能高。这个寻找"信息量最大"特征的过程一直完成，直到我们给定一个停止标准，在最终的叶节点结束。为此，我们必须选择一个合适的样本纯度指标，并设定决策树的深度（层数）和叶节点的最小样本点数。本研究选择基尼系数作为数据样本的度量标准。决策树的最大层深设定为 3~5，叶子样本的最小数量为 10~200。

对于本研究所涉及的探伤结果二分类问题，可将数据中探伤结果样例类别与分类器预测结果类别的组合划分为真正例(实际探伤未通过且预测正确)、假正例(实际探伤通过但预测错误)、真反例(实际探伤通过且预测正确)、假反例(实际探伤未通过但预测错误)四种情况，令 *TP*、*FP*、*TN*、*FN* 分别表示其对应数量。绘制 *P-R* 曲线用以说明模型预测效果，其中查准率 *P(precision)*与查全率 *R(recall)*，分别定义为：

$$P=TP/(TP+FP) \quad (1)$$

$$R=TP/(TP+FN) \quad (2)$$

ROC 曲线的纵轴是"真正例率"*(True Positive Rate，简称 TPR)*，横轴是"假正例率"*(False Positive Rate，简称 FPR)*，两者定义为：

$$TPR=TP/(TP+FN) \quad (3)$$

$$FPR=FP/(TN+FP) \quad (4)$$

模型预测效果的评估是通过 *AUC* 来实现的，*AUC* 是 *ROC* 曲线下面积分得到的[10]。*AUC* 显示了模型的预测能力，而训练集和测试集之间的 *AUC* 差值则显示了模型的泛化能力。*AUC* 值越大，模型的预测能力越强。训练集和测试集之间的 *AUC* 差值越小，模型的泛化能力越强。

1.2　模型优化

设定决策树模型最大深度分别为三层、四层和五层，通过计算 AUC 值来检验模型的预测能力，同时将叶子样本的最小值设定为 10 到 220。图 2 展示了不同最大层深下决策树模型中最小叶片样本数的变化对模型训练集和测试集 AUC 值的影响。如图 2(a)所示，训练集的 AUC 值高于测试集。随着叶片样本最小值的增加，训练集的 AUC 值略有下降。相反，当模型的最大深度为五层时，训练集的 AUC 值更高。另一方面，当最小叶片样本值设置为 40 时，测试集的 AUC 值逐渐增加。AUC 在初始阶段后保持较高水平，随着最小叶片样本数的增加而逐渐降低。相反，当模型的最大深度为四层、最小叶片样本数为 90 时，训练集的 AUC 最高。图 2(b)展示了三种最大层深度下训练集和测试集的 AUC 差值。当叶子样本的最小数量相对较少时，在训练数据集上训练的模型容易过拟合，导致测试集的 AUC 值较低，模型的泛化能力下降。只有当最小叶片样本数超过 70 时，训练集和测试集的 AUC 差值才会降到较低水平，表明泛化能力更强。可以确定模型模拟。结果表明，该模型在最大深度为 4 层和最小叶片样本数为 90 时预测能力和泛化能力的综合表现最佳。

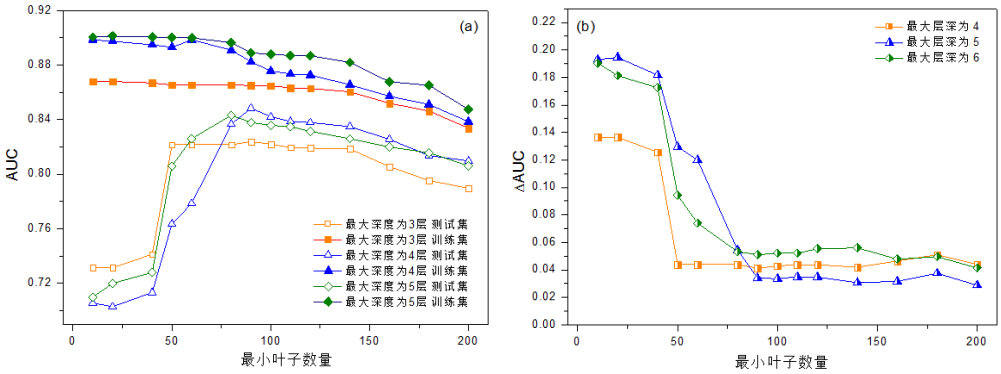


图 2　决策树结构参数对预测模型 AUC 的影响

Fig. 2　The Influence of Decision Tree Structure Parameters on AUC of Predictive Model

图 3 显示了决策树模型的精度/召回曲线。从图中可以看出，精确度随着召回率的增加而降低。在进行预测之前必须设置一个预测阈值，模型生成预测分析值，然后与阈值进行比较，最后做出决定。通常情况下，将阈值设置得高（接近 1）会提高精确度，而将其设置得低（接近 0）则会降低精确度。策略 1：为确保召回值，应提高合格板坯的阈值，这将提高召回率。但是，这种方法需要更充足的检测能力。因此，只有在有足够的检测能力的情况下，才能实施这一预测策略。策略 2：为优化检测资源，建议综合评估召回率和精确率。在较小范围内，尽可能多的检测出不符合质量要求的板坯。

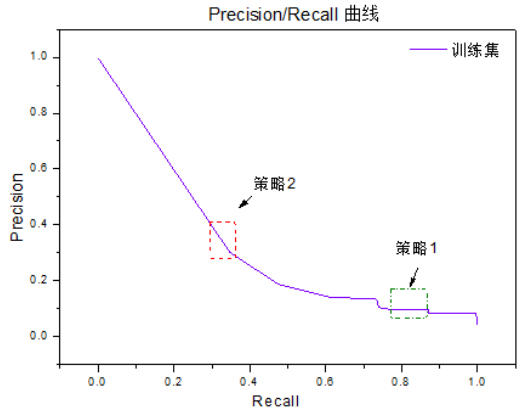
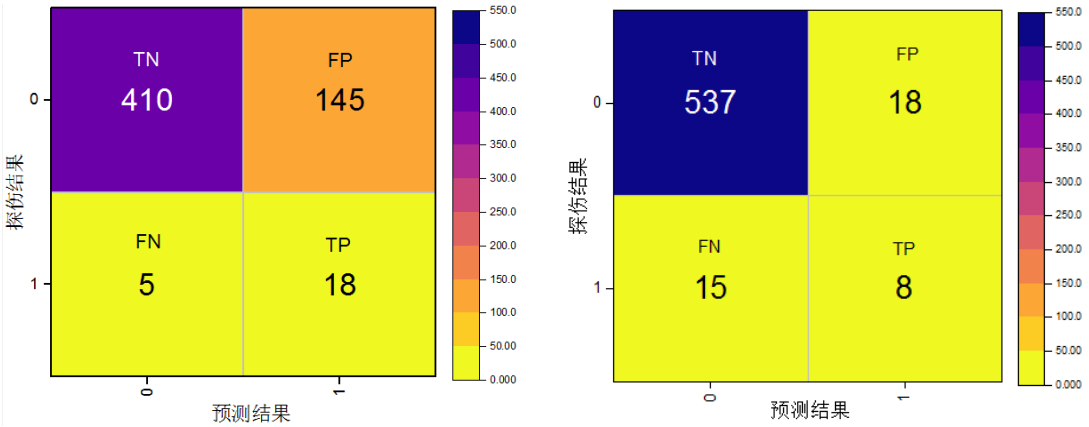


图 3　决策树模型的 P/R 曲线

Fig. 3　The P/R curve of the decision tree model

为了更清晰地展示使用上述两种策略对测试集进行预测的结果，我们在图 4 中列出了两种策略对应的混淆矩阵。图 4(a)显示了策略 1 引导下的测试集混淆矩阵。预测阈值设定为 0.85，召回率为 0.78，查准率为 0.12。图 4(b)显示了使用策略 2 时测试集的混淆矩阵。预测阈值设为 0.35，结果召回率为 0.347，查准率为 0.307。



(a) 预测阈值设定为 0.85　　(b) 预测阈值设定为 0.35

图 4　两种不同策略指导下的测试集混淆矩阵

Fig. 4　The Confusion matrix for the test set guided by different strategies

2　分析与讨论

2.1　决策树模型的可视化

在对模型结构与设置进行优化后，利用 Graphviz 模块对决策树模型进行可视化。决策树模型中每个节点的决定因素包括钢中 T.Al 含量、RH 净循环时间、浇铸过热度、连铸拉坯速度和轻压下轧辊间隙合格率。图 5 显示了预测管线钢连铸板坯探伤的决策树模型。该模型由四层组成，共有 10 个节点，端点处有 11 个叶子节点。如图所示，浇铸速度是影响管线钢连铸板坯探伤结果的最关键因素，它存在于决策树的各个层次。在实际生产中，通常是通过调整浇铸速度来匹配和调节生产节奏以及中间包温度的变化。这种调节具有明显的效果，但连铸过程中拉坯速度的波动会严重影响钢坯的冶金质量。拉速的变化会影响钢水的各个方面，包括动量传递、热量传递和质量传递[11]。这些因素会影响钢水流动、液穴形态、结晶器壁冷却强度、浮晶沉降、气体和夹杂物上浮以及耐火材料的溶蚀。因此，应严格控制连铸拉速的波动，并设定波动范围将有效提高铸坯和钢板的冶金质量。

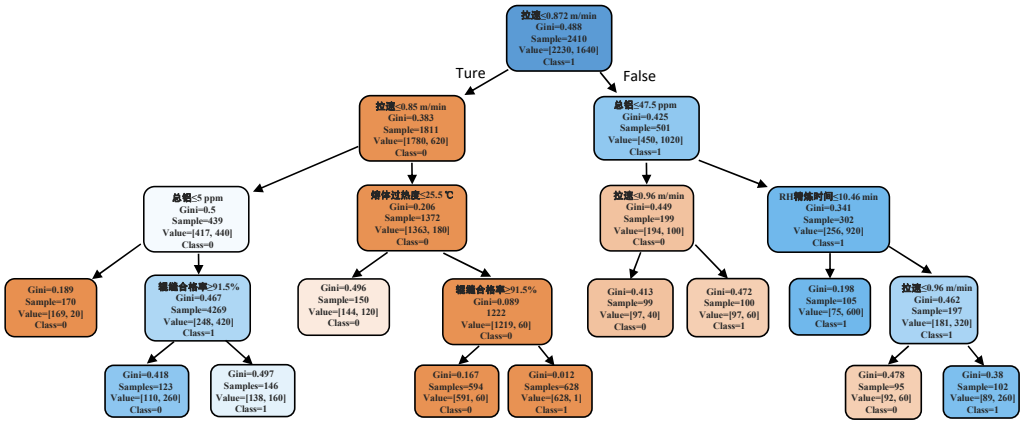


图 5　连铸板坯探伤预测的可视化决策树

Fig. 5　Visualization of the decision tree prediction model

2.2　重要性指数分析

为了定量地描述每个特征属性对管线钢连铸坯探伤结果的影响，本研究利用机器学习算法中的重要性指数来评估每个关键工艺因素的重要性。如图 6 所示，说明了比较各关键工艺因素对缺陷预测结果影响的重要性。连铸牵引速度的重要性指数为 61.8%，排名第一。铸造过热的重要性指数为 20.48%，排名第二。钢液中总铝含量、RH 净循环时间和辊缝合格率的重要指数分别为 8.63%、6.27%和 2.72%。结合决策树的可视化和各关键工艺因素重要性百分比的比较结果，可以得出以下工艺控制注意事项：

(1) 铸造速度对钢板探伤的最终结果有重要影响。因此，应严格控制该工艺参数，不得超过 0.87m/min；

(2) 浇铸过热度（铸造温度）会直接影响熔流的粘度。如果浇铸温度过低，将不利于气体和夹杂物的去除。因此，熔体的过热温度应控制在 25.5℃以下；

(3) RH 精炼过程中合理的精炼时间对于有效的熔体脱气和完全去除夹杂物至关重要。根据决策树分类法，建议 RH 的最短精炼时间为 10.5 分钟。

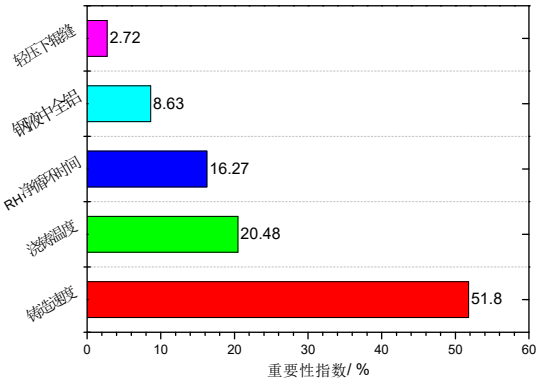


图 6　各关键工艺因素对探伤结果影响的重要性比较

Fig. 6　Comparison of the importance of key process factors

3 结论

影响管线钢连铸板坯质量因素较多，如各元素含量、RH 处理周期、钢中总铝含量、RH 净循环时间、钢液浇铸过热度、拉速、辊缝合格率等，这些因素等都会对最终钢板探伤结果产生不同的影响。众多影响因素与探伤结果存在复杂的非线性关系，传统的基于数理统计与冶金学原理的实验对探伤结果预测与分析具有一定的局限性。通过本研究决策树预测模型的开发与评估得到的以下三点结论：

（1） 本研究基于工业生产大数据，以 RH 精炼环节与连铸环节各项关键工艺点为特征属性，采用 CART 分类决策树算法建立管线钢连铸板坯探伤预测模型，可实现对管线钢连铸板坯探伤结果的高精度预测。

（2） 经过参数调优后的决策树算法模型通过工作特征曲线（ROC）评价预测证实对于正、反例的预测都具有较高的水平，且具有很强泛化能力。

（3） 通过决策树可视化与重要性指标比对定量地分析了给定样本范围内各关键工艺参数对管线钢连铸板坯探伤预测结果的影响，为探伤结果预测与工艺控制提供可靠依据。

参考文献：

[1] Wang XD, Yan YJ. Promoting the technical progress of steel industry with intelligent manufacturing [J].Metallurgical Industry Automation, 2019, 43(1): 1.

[2] Nkonyana T, Sun Y, Twala B, et al. Performance Evaluation of Data Mining Techniques in Steel Manufacturing Industry [J]. Procedia Manufacturing, 2019, 35.

[3] Mingji Liu, Wenzhao Li. Prediction and Analysis of Corrosion Rate of 3C Steel Using Interpretable Machine Learning Methods [J], Materials Today Communications,2023, 106408.

[4] Dong G B，Li X C，Zhao J X，et al. Machine learning guided methods in building chemical composition-hardenability model for wear-resistant steel [J]. Materials Today Communication，2020(24), 101332.

[5] Duc-Kien Thai, Dai-Nhan Le, Quoc Hoan Doan, Thai-Hoan Pham, Dang-Nguyen Nguyen, Classification models for impact damage of fiber reinforced concrete panels using Tree-based learning algorithms, Structures, Volume 53, 2023, Pages 119-131.

[6] Yimian Chen, Shuize Wang, Jie Xiong, Guilin Wu, Junheng Gao, Yuan Wu, Guoqiang Ma, Hong-Hui Wu, Xinping Mao, Identifying facile material descriptors for Charpy impact toughness in low-alloy steel via machine learning, Journal of Materials Science & Technology, Volume 132, 2023,Pages 213-222.

[7] Chunyuan Cui, Guangming Cao, Xin Li, Zhiwei Gao, Jianjun Liu, Zhenyu Liu, A strategy combining machine learning and physical metallurgical principles to predict mechanical properties for hot rolled Ti micro-alloyed steels, Journal of Materials Processing Technology, Volume 311, 2023, 117810.

[8] Fan Xia, Zhiwei Li, Ming Ma, Yonggang Zhao, Changjun Wu, Xuping Su, Haoping Peng, Effect of Nb on microstructure and corrosion resistance of X80 pipeline steel, International Journal of Pressure Vessels and Piping, Volume 203, 2023, 104949.

[9] Saman Nadizadeh Shorabeh, Najmeh Neysani Samany, Foad Minaei, Hamzeh Karimi Firozjaei, Mehdi Homaee, Ali Darvishi Boloorani, A decision model based on decision tree and particle swarm optimization algorithms to identify optimal locations for solar power plants construction in Iran, Renewable Energy, Volume 187, 2022, Pages 56-67.

[10] Tom Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, Volume 27, Issue 8, 2006, Pages 861-874.

[11] Jie Yang, Dengfu Chen, Mujun Long, Huamei Duan, Transient flow and mold flux behavior during ultra-high speed continuous casting of billet, Journal of Materials Research and Technology, Volume 9, Issue 3, 2020, Pages 3984-3993.