

# 基于随机森林的热轧带钢质量分析与预测方法

纪英俊<sup>1</sup>, 勇晓玥<sup>1</sup>, 刘英林<sup>2</sup>, 刘士新<sup>1</sup>  
(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110819; 2. 上海宝信软件股份有限公司 大数据事业部, 上海 201203)

**摘 要:** 以某钢铁企业的热轧带钢生产实际数据作为分析对象, 基于改进的随机森林算法分析工艺参数与产品质量间的隐含关系, 进行影响产品质量关键工艺参数的特征提取, 建立热轧带钢产品缺陷预测模型. 实验结果表明, 对非平衡数据集进行平衡处理可以提高样本预测精度; 采用 CART 与 C4.5 相结合的方法比单一方法可以进一步提升预测精度; 同时根据特征的高相关与低相关特性, 将互信息作为评价指标应用于特征选择, 可以提升随机森林算法的分类效果. 在以上三种改进策略下, 热轧带钢缺陷的识别率得到明显提高.  
**关 键 词:** 热轧带钢; 缺陷预测; 数据驱动; 特征提取; 随机森林  
**中图分类号:** TP 277      **文献标志码:** A      **文章编号:** 1005-3026(2019)01-0011-05

## Random Forest Based Quality Analysis and Prediction Method for Hot-Rolled Strip

Ji Ying-jun<sup>1</sup>, YONG Xiao-yue<sup>1</sup>, LIU Ying-lin<sup>2</sup>, LIU Shi-xin<sup>1</sup>  
(1. School of Information Science & Engineering, Northeastern University, Shenyang 110819, China; 2. Big Data Department, Shanghai Baosight Software Co., Ltd., Shanghai 201203, China. Corresponding author: LIU Shi-xin, E-mail: sxliu@mail.neu.edu.cn)

**Abstract:** The process data of hot-rolled strips from an iron and steel enterprise were analyzed to find out the inherent relationship between process parameters and production quality by using an improved random forests algorithm. After critical features being extracted, a defect prediction model was built. According to the experiment, balancing operation can improve the prediction accuracy of the imbalanced data sets. Meanwhile, the combination of CART and C4.5 can further improve the prediction accuracy than each single method. Furthermore, in consideration of the characteristics whose features have high or low correlations with the response variable, mutual information was introduced as an evaluation criterion for feature selection. Mutual information makes great contribution to classification effect of random forest algorithm, and recognition rate of defects of hot-rolled strips is obviously improved by using three strategies.  
**Key words:** hot-rolled strip; defect prediction; data driven; feature selection; random forests

热轧带钢产品质量是决定钢铁企业制造成本及服务水平的重要因素, 一直受到业界和学术界的关注和研究. 热轧带钢产品质量指标主要包括表面质量、力学性能、尺寸精度<sup>[1]</sup>. 在表面缺陷检测研究方面, Ghorai 等<sup>[2]</sup>开发了集成钢厂自动化视觉检测系统, 使用支持向量机智能识别热轧钢表面缺陷. 在力学性能预测研究方面, Sui 等<sup>[3]</sup>针对具有高维、强耦合和冗余特征的热轧工艺质量参数, 先采用 Gram-Schmidt 正交变换组合信息熵方法来选择特征子集, 然后采用极限学习机建立预测模型, 实现了对带钢力学性能的预测. 在尺寸精度控制研究方面, Li 等<sup>[4]</sup>在考虑影响轧制间隙精度的多种因素下, 提出了基于支持向量机的回归模型, 预测轧制间隙, 提高了带钢厚度控制精度. 综上所述, 机器学习方法与带钢质量问题已有广泛结合, 可以应用于多种参数的预测及辅助钢铁生产过程决策等方面. 在实际应用中, 关键问题在于如何根据数据特点选择匹配的机器学习方法

来完成缺陷的分析与预测。

本文根据某钢铁企业热轧带钢的实际生产数据,首先对数据进行预处理得到初始样本,再通过特征选择对包含大量冗余、强耦合特征的高维数据进行降维,保证模型的精度和可解释性.之后针对非平衡数据集的二分类问题,采用改进的随机森林算法,建立解决热轧带钢缺陷识别问题的随机森林模型.最后通过 K 折交叉验证来验证分类模型的精度,并结合混淆矩阵、ROC 曲线等指标评判分类结果.

# 1 改进的随机森林算法

## 1.1 数据预处理与特征提取

本文针对带钢表面挂腊-辊印缺陷进行研究.原始数据包含 407 个生产工艺参数特征,2 278 个样本,其中存在目标缺陷样本 24 个,其余样本未发生缺陷.样本数据表明该问题属于非平衡数据集下的二分类问题.

采用多重插补法<sup>[5]</sup>对原始数据进行缺失值、异常值预处理后,根据特征的特点依次采用后向逐步选择、群集选择与最优子集选择进行特征提取,筛选时采用最优调整  $R^2$  值与  $p$  值作为评价指标,最终选出 11 个特征变量用于模型建立,特征信息如表 1 所示.

表 1 最优子集选择后的特征			
Table 1 Features after best subset selection			
编号	特征名称	编号	特征名称
1	精轧入口实际温度	7	F2 轧制力实际值
2	板坯质量	8	卷取目标温度
3	F3 轧机力矩实际值	9	终轧目标温度
4	F1 轧机力矩实际值	10	F4 轧机力矩实际值
5	F3 轧制力设定值	11	卷筒最小张力设定值
6	F6 轧制力设定值		

## 1.2 基于 NCL 和 SMOTE 混合的非平衡数据集改进方法

本文样本数据中包含合格品与缺陷品比例为 18:1,这种数据的不平衡性会造成少数类难以识别,预测结果偏向多数类,分类器的分类精度下降.因此,本文采用近邻消除法(neighborhood cleaning rule, NCL)<sup>[6]</sup>与合成少数类过采样技术(synthetic minority over-sampling technique, SMOTE)<sup>[7]</sup>相结合的方法来改善数据集的不平衡性.本文将 NCL 和 SMOTE 算法优点相结合,提出 NCL 和 SMOTE 混合方法.首先设定一个数据

集平衡比例,利用 NCL 算法去除多数类中的噪声样本,再用 SMOTE 算法人工合成少数类样本,循环迭代直至达到数据集平衡比例后,跳出循环.图 1 是改进的 NCL + SMOTE 算法流程.

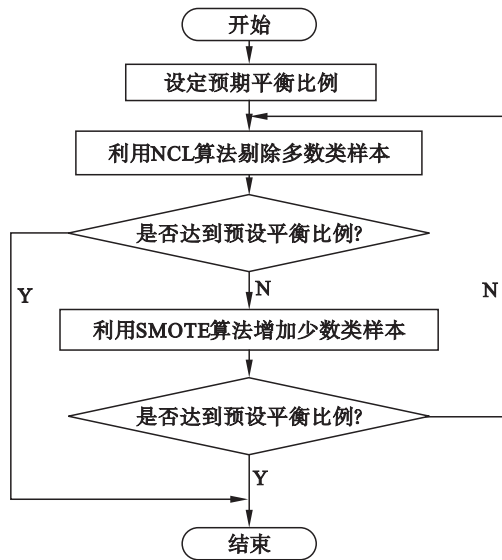


图 1 NCL + SMOTE 算法流程  
Fig. 1 Flowchart of NCL-SMOTE algorithm

## 1.3 决策树分裂节点选择算法

对于随机森林算法,树的分裂节点选择算法会直接决定随机森林的分类效果. CART 算法<sup>[8]</sup>和 C4.5 算法<sup>[9]</sup>都是经典且有效的分类节点选择算法,但两者评价指标的差异会导致最后生成决策树的差异,分类结果也就各不相同.本文提出 CART 与 C4.5 混合算法,将信息增益与 Gini 系数两种评价指标相结合,发挥两者的核心优势,以此来提高决策树的分类精度.

混合算法评价指标 mix 计算公式如下:设样本训练集为  $D$ ,样本中的特征为  $A$ ,

$$\text{mix} = \alpha_1 \text{Gini}_{\text{split}}(D) - \alpha_2 \text{GainRatio}(D, A). \quad (1)$$
其中:  $\alpha_1, \alpha_2 \in [0, 1]$ , 且两者不能同时为 0 或 1;  $\text{GainRatio}(D, A)$  为信息增益率;  $\text{Gini}_{\text{split}}(D)$  为 Gini 系数,计算方法与 CART 和 C4.5 算法下给出方式相同.混合算法中拥有最小 mix 值的特征为当前条件下的最优分裂节点.

## 1.4 随机森林的特征选择

决策树的分类精度<sup>[10]</sup>和树间相似度是直接影响随机森林分类效果的关键因素.在构建森林时涉及特征选择,若要保证树的分类精度,特征集中需包含与目标变量高度相关的特征,起分类主导作用,称其为高相关特征.同时需在特征集内选取一部分低相关特征保持树间差异,提升泛化能力,称为低相关特征.本文以互信息<sup>[11]</sup>作为评价指标来衡量特征与目标变量间的相关度.

设两个离散值变量  $X, Y$ , 它们的互信息如式 (2) 所示, 单位为 bit.

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \lg \left( \frac{p(x, y)}{p(x)p(y)} \right).$$

(2)

其中:  $p(x, y)$  是  $X$  和  $Y$  的联合概率分布;  $p(x)$  和  $p(y)$  是  $X$  和  $Y$  的边缘概率分布. 当变量中属性都为连续值时,  $X$  和  $Y$  的互信息为

$$I(X, Y) = \int_Y \int_X p(x, y) \lg \left( \frac{p(x, y)}{p(x)p(y)} \right) dx dy.$$

(3)

在构建特征子集时, 首先利用互信息将特征集划分为高相关区和低相关区, 而后按照预设的抽取比例, 从两区域各随机抽取一部分特征构成决策树的特征子集.

1.5 改进的随机森林算法

基于以上 3 种改进策略, 本文提出的改进随机森林算法计算流程如下.

步骤 1 计算每个特征的互信息值, 而后对其进行从高到低排序.

步骤 2 设定高相关和低相关特征数量比例为  $1:p$ , 再根据比例划分为高相关区与低相关区.

步骤 3 设森林中有  $B$  棵决策树, 对于每一棵决策树有:

- 1) 按照 Bagging 方法从原数据集中抽取样本子集;
- 2) 从高相关区抽取  $\frac{k}{1+p}$  个特征, 并从低相关区抽取  $\frac{p}{1+p}k$  个特征, 组成特征子集;

3) 利用样本子集和特征子集构建决策树, 并加入到森林中.

步骤 4 森林构建完成.

采用排序后再分区抽取的方法, 可以同时保证树的分类精度与树间差异性, 利于提升随机森林的分类效果.

2 实验结果

为了分析三种随机森林算法改进策略的有效性, 本文基于相同的预处理数据进行了 3 组实验. 实验采用 5 折交叉验证法<sup>[12]</sup>, 设置随机森林中树的数量为 100, 结果采用敏感度、特异度、准确度、几何平均数  $G_{\text{mean}}$  和曲线下面积 (area under curve, AUC) 作为评价指标.

2.1 非平衡数据改进结果

对非平衡数据集下的训练集进行平衡处理到指定比例, 根据不同的平衡比例进行了 5 组实验, 结果如表 2 所示.

比较原始数据集, 5 种不同比例的改进数据集对板坯是否产生缺陷的分类效果都有所提升. 采用 NCL + SMOTE 算法将正负样本比例调节到 1.1:1 时, 可以发现其分类效果尤其在负样本的命中率方面, 牺牲小部分敏感度换取了特异度的明显提升,  $G_{\text{mean}}$  值有较大提升. 通过实验发现 SMOTE 算法可以为平衡比例的选取提供参考, 其对负样本的命中率有较大提升, 且与 NCL 算法相结合可以很好地减少 NCL 算法的时间复杂度, 提升分类效果.

表 2 非平衡数据集改进结果  
Table 2 Evaluation results of improved imbalanced data sets

算法	正负样本比例	混淆矩阵	敏感度	特异度	$G_{\text{mean}}$	AUC
未平衡处理	18. 2:1	$\begin{bmatrix} 434 & 2 \\ 16 & 8 \end{bmatrix}$	0. 995 4	0. 333 3	0. 655 0	0. 81
NCL	17. 2:1	$\begin{bmatrix} 436 & 0 \\ 14 & 10 \end{bmatrix}$	1	0. 416 6	0. 645 5	0. 82
SMOTE	4. 26:1	$\begin{bmatrix} 428 & 8 \\ 9 & 15 \end{bmatrix}$	0. 981 6	0. 625	0. 783 3	0. 82
SMOTE	1. 15:1	$\begin{bmatrix} 424 & 12 \\ 9 & 15 \end{bmatrix}$	0. 972 4	0. 625	0. 779 6	0. 82
NCL + SMOTE	4. 18:1	$\begin{bmatrix} 428 & 8 \\ 10 & 14 \end{bmatrix}$	0. 981 6	0. 583 3	0. 756 7	0. 83
NCL + SMOTE	1. 1:1	$\begin{bmatrix} 424 & 12 \\ 8 & 16 \end{bmatrix}$	0. 972 4	0. 666 6	0. 805 2	0. 83

2.2 决策树分裂节点算法改进结果

本实验采用 2.1 节正负样本比例为 1.1:1 的数据集作为训练集, 通过实验对比该数据集下采用 C4.5, CART 以及 C4.5 与 CART 相加权的方

法这 3 种分裂树节点算法的分类效果. 本文把这种通过不同权重配比寻找最优组合策略的加权方法称为 ACC 混合方法. 该方法中的两个参数为上述两种基础方法的权重系数. 为了克服随机性, 所

有实验结果取 100 次求解后的平均值. 表 3 为 CART, C4.5 及两者混合算法实验对比与评价结

表 3 CART, C4.5 及两者混合的决策树分裂节点算法对比与评价结果  
Table 3 Evaluation and comparison results of split node algorithms CART, C4.5 and hybrid method

算法	混淆矩阵	敏感度	特异度	$G_{\text{mean}}$
CART	$\begin{bmatrix} 424 & 12 \\ 8 & 16 \end{bmatrix}$	0.972 4	0.666 6	0.805 2
C4.5	$\begin{bmatrix} 418 & 18 \\ 8 & 16 \end{bmatrix}$	0.958 7	0.666 6	0.799 5
ACC(0.9,0.1)	$\begin{bmatrix} 424 & 12 \\ 8 & 16 \end{bmatrix}$	0.972 4	0.666 6	0.805 2
ACC(0.8,0.2)	$\begin{bmatrix} 426 & 10 \\ 8 & 16 \end{bmatrix}$	0.977 0	0.666 6	0.807 1
ACC(0.7,0.3)	$\begin{bmatrix} 426 & 10 \\ 7 & 17 \end{bmatrix}$	0.977 0	0.708 3	0.831 9
ACC(0.6,0.4)	$\begin{bmatrix} 424 & 12 \\ 6 & 18 \end{bmatrix}$	0.972 4	0.75	0.854 0
ACC(0.5,0.5)	$\begin{bmatrix} 420 & 16 \\ 6 & 18 \end{bmatrix}$	0.963 3	0.75	0.849 9
ACC(0.4,0.6)	$\begin{bmatrix} 423 & 13 \\ 7 & 17 \end{bmatrix}$	0.970 1	0.708 3	0.828 9
ACC(0.3,0.7)	$\begin{bmatrix} 422 & 14 \\ 8 & 16 \end{bmatrix}$	0.967 8	0.666 7	0.803 2
ACC(0.2,0.8)	$\begin{bmatrix} 420 & 16 \\ 8 & 16 \end{bmatrix}$	0.963 3	0.666 7	0.801 4
ACC(0.1,0.9)	$\begin{bmatrix} 419 & 17 \\ 8 & 16 \end{bmatrix}$	0.961 0	0.666 7	0.800 4

从表 3 的结果可以看出,ACC 混合算法的特异度和敏感度要高于 C4.5 算法与 CART 算法. 从  $G_{\text{mean}}$  值也可以看出,ACC 混合算法会获得更好效果. 当  $\alpha_1=0.6, \alpha_2=0.4$  时算法的效果最佳, 此时两种算法的权重达到此数据集下的最佳配比. 与单独使用 CART 和 C4.5 算法的分类结果比较, ACC 混合算法下负样本命中率提升了 8.3% .

2.3 改进的随机森林特征选择算法结果

在 2.1 节与 2.2 节给出的最优条件下,进行随机森林特征选择改进算法实验. 实验中特征的选取由随机选择变为按相关度高低分区选择, 并采用三种不同方法(Log, Sqrt 和 None) 确定决策

树所需特征数目,得到了按高低分区排序方法与未进行高低分区排序方法的对比结果,如表 4 所示.

从结果可以看出,采用高低分区排序的特征子集选择方法的结果都明显优于未排序的随机选择结果. 由于原数据集特征数为 11 个, 使得 Log 和 Sqrt 特征抽取方法下恰巧得到了相同结果. 最终正样本命中率为 96.7%, 负样本命中率为 83.3%, 总体命中率达到 96.1%, 置信度 0.92, 预测结果可信度较高. 对特征进行相关度分区后,可以保证森林内树的强度与树间差异性,提高随机森林算法的分类效果. 从图 2 可以明显看出,改进的随机森林算法有更好的 ROC 曲线, AUC 值提

表 4 改进的特征选择算法实验结果  
Table 4 Evaluation results of improved feature selection algorithm

方法	混淆矩阵	敏感度	特异度	$G_{\text{mean}}$	AUC
Log	$\begin{bmatrix} 422 & 14 \\ 4 & 20 \end{bmatrix}$	0.967 8	0.833 3	0.898 1	0.92
Sqrt	$\begin{bmatrix} 422 & 14 \\ 4 & 20 \end{bmatrix}$	0.967 8	0.833 3	0.898 1	0.92
None	$\begin{bmatrix} 411 & 25 \\ 6 & 18 \end{bmatrix}$	0.942 6	0.75	0.840 8	0.90



升 13.6% ,相比传统随机森林算法在分类效果上有显著提高.

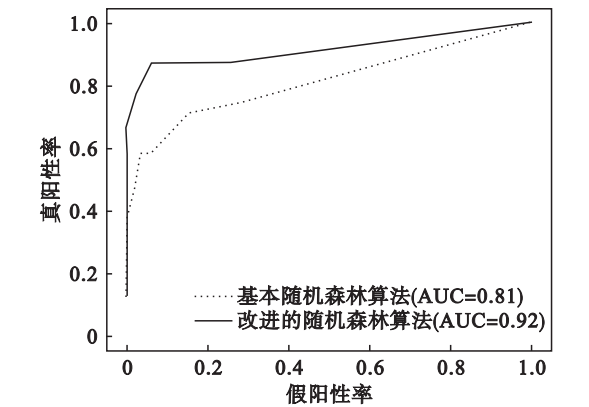


图 2 两种算法的 ROC 对比曲线

Fig. 2 Comparison of ROC curves of two algorithms

### 3 结 语

基于热轧带钢生产工艺实际数据,采用改进的随机森林算法作为数据分析方法,对热轧带钢的缺陷识别问题进行了分析,应用本文改进的随机森林算法,缺陷带钢的识别率得到了明显提高.实验结果表明,本文改进的随机森林算法可以有效地解决热轧带钢产品缺陷预测问题.

本文采用随机森林算法作为基础分类算法,优化对象为随机森林算法中的分裂节点和特征选择方法,采用互信息来判定特征间的重要程度是一种有效的方法,但如何结合随机森林特点,深度优化特征子集选择过程,选出最有效特征,仍是提高分类精度的关键工作,需要未来进一步研究.

#### 参考文献：

[ 1 ] 王永胜,成泽伟,李宏,等. 热轧板坯表面缺陷分析[ J ]. 钢铁研究学报,2002,14(2) :75 – 76.

( Wang Yong-sheng, Cheng Ze-wei, Li Hong, et al. Analysis on surface defect of hot rolled slab[ J ]. *Journal of Iron and Steel Research*,2002,14(2) :75 – 76. )

[ 2 ] Ghorai S, Mukherjee A, Gangdaran M, et al. Automatic defect detection on hot-rolled flat steel products [ J ]. *IEEE Transactions on Instrumentation and Measurement*,2013,63(3) :612 – 621.

[ 3 ] Sui X Y, Lyu Z M. Prediction of the mechanical properties of hot rolling products by using attribute reduction ELM[ J ]. *International Journal of Advanced Manufacturing Technology*,2016,85(5/6/7/8) :1395 – 1403.

[ 4 ] Li W, Yao X L, Yu L, et al. Application of SVM regression in HAGC system [ C ]//The 27th Chinese Control and Decision Conference( CCDC). Qingdao,2015;3490 – 3494.

[ 5 ] Royston P. Multiple imputation of missing value[ J ]. *Stata Journal*,2004,4(3) :227 – 241.

[ 6 ] Laurikkala J. Instance-based data reduction for improved identification of difficult small classes[ J ]. *Intelligent Data Analysis*,2002,6(4) :311 – 322.

[ 7 ] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[ J ]. *Journal of Artificial Intelligence Research*,2002,16(1) :321 – 357.

[ 8 ] Leo B. Classification and regression trees[ M ]. Monterey: Cole Publishing,1984.

[ 9 ] Quinlan J R. C4. 5: programs for machine learning[ M ]. San Francisco: Morgan Kaufmann Publishers,1994.

[ 10 ] 雍凯. 随机森林的特征选择和模型优化算法研究[ D ]. 哈尔滨: 哈尔滨工业大学,2008.

( Yong Kai. Research on feature selection and model optimization of random forest[ D ]. Harbin: Harbin Institute of Technology,2008. )

[ 11 ] Peng H C, Long F, Ding C. Feature selection based on mutual information; criteria of max-dependency, max-relevance, and min-redundancy[ J ]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,2005,27(8) :1226 – 1238.

[ 12 ] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[ C ]// Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. Montreal,1995;1137 – 1143.