文章编号：2095-2163(2020)12-0145-05 中图分类号：TG142;TP181 文献标志码：A

基于改进集成学习算法的钢材质量预测

梁博德，王瑞敏，孙践知
（北京工商大学 计算机学院，北京 100048）

摘 要：钢材表面质量不合格会降低钢材的强度，甚至发生安全事故。为了对钢材质量进行在线诊断，本文利用采集的钢材实际生产数据建立了钢材质量预测模型，在采用特征提取算法找出影响钢材质量的关键工艺参数的基础上，通过分析工艺参数和钢材质量间潜在的关系，提出了一种适用于钢材质量预测的倾向性异质装袋集成学习算法。该集成学习算法用异质装袋方法生成个体分类器，并采用带有倾向性的结合策略。实验结果表明，改进方案可有效提升集成模型的分类性能以及稳定性，在钢材质量预测问题上具有较高的精度，模型预测准确率的均值为 91.4%，方差为 5.1%，可应用于实际生产。

关键词：钢材质量预测；特征提取；集成学习

Steel quality prediction based on improved ensemble learning algorithm

LIANG Bode, WANG Ruimin, SUN Jianzhi
(College of Computer Science, Beijing Technology and Business University, Beijing 100048. China)

【Abstract】 Steel surface quality is not qualified will reduce the strength of steel, and even safety accidents. In order to diagnose the steel quality online, a steel quality prediction model was established based on the actual production data collected. Based on the feature extraction algorithm to find out the key process parameters affecting the quality of steel, and by analyzing the potential relationship between the process parameters and the quality of steel, an integrated learning algorithm for predicting the quality of steel was proposed. The ensemble learning algorithm uses heterogeneous bagging method to generate individual classifiers and adopts a binding strategy with bias. The experimental results show that the improved scheme can effectively improve the classification performance and stability of the integrated model, and has a high accuracy in steel quality prediction. The mean prediction accuracy of the model is 91.4%, and the variance is 5.1%, which can be applied to actual production.

【Key words】 Steel quality prediction; Feature selection; Ensemble learning

0 引 言

钢材质量指标主要分为表面质量、力学性能、尺寸精度 3 个方面[1]，其质量主要由成分、结构、制备工艺等因素决定，其中某个环节的改变都可能对最终钢材的质量产生巨大的影响[2]。

钢材生产的过程连续且繁杂，采集到的数据往往具有高冗余、高噪声、低精度、强耦合等特点，但通过数据清洗、特征提取以及挑选适当的机器学习方法后，可以挖掘出数据中本质的规律。

在相关研究中，纪英俊等针对实际的钢材生产数据集，在使用了 SMOTE 算法的平衡处理基础上，再使用 CART 与 C4.5 相结合的随机森林的方法，分析工艺参数与产品质量间的隐含关系，并提取出了影响产品质量的关键工艺参数[3]；Sui 等针对具有高维度、强耦合和冗余信息等特点的热轧生产过程工艺参数，提出了一种基于 ELM 算法的模型与特制提取方法相结合的新的力学性能预测模型[4]；熊鹰等将计算机视觉应用于表面缺陷识别中，取得了较高的识别率[5]；杨威等使用随机森林算法以及数据与机理分析相结合的力学性能建模方法，建立了具有较高精度的热轧带钢力学性能预报模型，并筛选出性能预报模型的影响因素[6]；吴思炜等采用了马氏距离剔除异常值，使用平均影响值法筛选出对力学性能影响较大的工艺参数，再结合贝叶斯神经网络算法，建立了精度较高的钢材力学性能预报模型[7]。

其它工业产品质量监控领域中，Chien 等使用 K-Means 和决策树等方法对半导体的制造数据进行了研究，推断可能的故障原因和制造工艺的变化，并应用在实际生产中，提升了产品合格率[8]；皮骏等将基于遗传算法优化的支持向量机（GA-SVM）成功应用于航空发动机磨损和刀具的故障诊断上，并与

基金项目：2018 年工业互联网创新发展工程（2018282-41）。

作者简介：梁博德（1994-），男，硕士研究生，主要研究方向：工业大数据、集成学习；王瑞敏（1996-），男，硕士研究生，主要研究方向：工业大数据；孙践知（1967-），男，硕士，教授，主要研究方向：无线传感器网络、工业大数据。

通讯作者：王瑞敏 Email：1093861355@qq.com

收稿日期：2020-10-1

BP 神经网络的预测结果进行了对比,证明了在小样本数据集上 GA-SVM 算法的优越性[9]。江琨等使用 XGBoost 集成学习算法对工业产品质量进行了预测[10]。

本文根据某钢铁公司的实际生产数据,经数据预处理得到初始样本,再通过相关性分析、特征提取和主成分分析方法对数据进行了降维,去除初始样本的冗余性、强耦合性后,提出倾向性异质装袋算法(Propensity Heterogeneous Bagging,PHB)对钢材表面质量监控问题建立了分类模型。实验结果表明,本文的优化策略可以有效提升集成学习模型的分类性能,并且在钢材质量分类问题上具有较高的精度。

1 算法描述

1.1 数据预处理与特征提取

本文数据为某钢铁企业通过 MES 系统记录下来的一部分具有 S501 类型表面缺陷钢材的炼钢、热轧、退火、酸洗 4 个工艺流程中的实际生产数据,以及钢材在酸洗后的质检结果,工艺流程和数据采集过程如图 1 所示。

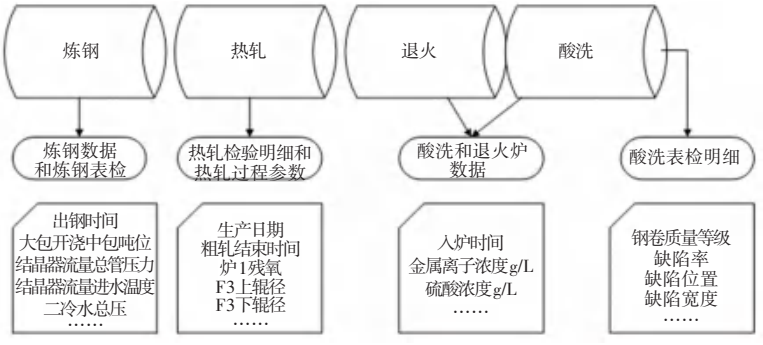

图 1 数据采集过程

Fig. 1 Data collection process

在智能制造建模过程中,需要将冗余和误差较大的数据剔除,保证训练数据和预测数据的均匀分布,才能建立出包含正确规律的模型[11]。本文采用 Filter 和 Embedding 方法相结合的方式进行特征选择,在此基础上对提取的特征使用主成分分析进行特征融合,去除了特征之间的高耦合性和冗余性,得到建立模型的最优数据集。

1.2 集成学习

集成学习是机器学习方法的一种,通过结合各种分类器以实现更好的预测性能。集成学习方法首先预测一些初步的结果,然后将初步结果组合生成一个新的最终结果[12]。研究表明,集成分类器通常比基础分类器表现得更好[13]。集成学习主要算法分为 Boosting,Bagging 和 Stacking。Bagging 是使用不同数据集产生基础分类器的方法,每个基础分类器都有自己的训练集,通常使用随机抽取方法和替换产生不同训练集,在生成所有训练集之后为每个分类器构建模型。最终的预测结果通过投票结合起来。这种集成学习方法减少了过拟合问题,并且对不稳定学习算法更有效,其中最著名的就是随机森林(Random Forest)算法。Boosting 在提高某些机器学习模型的预测精度方面也具有卓越的性能。在所有 Boosting 算法中,在每个学习阶段,实例都会重新加权,将在下一步中选择错误的分类实例的概率增加,以便模型在下次遇见时对它们进行正确分类,分类器的最终结果由多数投票组合[14]。Boosting 算法中最常用的为 AdaBoosting 算法。Stacking 是将分类器的预测结果作为新特征,并在新训练集上训练最终模型。

集成学习方法分两步构建:首先构建多个基础分类器,然后组合这些分类器。组合策略在基础分类器性能相差较大时,一般采用加权平均法,集成学习器的最终投票结果 y 和基础分类器的投票权重 ω_i 以及投票结果 a_i 的关系为式(1):

$$y = \sum \omega_i * a_i. \tag{1}$$

1.3 改进集成学习算法

在实际应用中,往往根据具体问题选择对应的算法。经仿真实验后发现,已有的基础算法和集成学习算法的分类准确率和模型稳定性均不够理想。本文在已有算法的基础上进行改进,以得到分类准确率更高、模型稳定性更好的模型。

1.3.1 集成学习优化理论

对于给定样本 x 和对应的回归任务 f,定义基础学习器 h_i 在训练样本 x 上的“分歧”A 为式(2):

$$A(h_i \mid x) = (h_i(x) - H(x))^2. \tag{2}$$

则基于加权平均策略的集成学习器 H 的“分

歧”$\bar{A}$为式(3)：

$$\bar{A}(h \mid x) = \sum_{i=1}^{T} \omega_i A(h_i \mid x) = \sum_{i=1}^{T} \omega_i \left(h(x)_i - H(x)\right)^2. \tag{3}$$

基础学习器 h_i 和集成 H 误差的平方分别为式(4)和式(5)：

$$E(h_i \mid x) = (f(x) - h_i(x))^2, \tag{4}$$

$$E(H \mid x) = (f(x) - H(x))^2, \tag{5}$$

令 $\bar{E}(h_i \mid x) = \sum_{i=1}^{T} \omega_i E(h_i \mid x)$ 表示基础学习器误差的加权平均值 $\bar{E}$，则有式(6)：

$$\bar{A}(h \mid x) = \sum_{i=1}^{T} \omega_i E(h_i \mid x) - E(H \mid x) = \bar{E}(h_i \mid x) - E(H \mid x). \tag{6}$$

令 $p(x)$ 表示样本的概率密度，则式(6)推广至全体样本上得式(7)：

$$\sum_{i=1}^{T} \omega_i \int A(h_i \mid x)\, p(x)\, dx = \sum_{i=1}^{T} \omega_i \int E(h_i \mid x)\, p(x)\, dx - \int E(H \mid x)\, p(x)\, dx. \tag{7}$$

基础学习器 h_i 在全体样本上的泛化误差和分歧项以及集成的泛化误差推广至全体样本时有式(8)~式(10)：

$$E(h_i) = \int E(h_i \mid x)\, p(x)\, dx, \tag{8}$$

$$A(h_i) = \int A(h_i \mid x)\, p(x)\, dx. \tag{9}$$

$$E(H) = \int E(H \mid x)\, p(x)\, dx. \tag{10}$$

将式(8)~(10)代入式(7)中，得出集成算法的泛化误差 E 为式(11)：

$$E = \bar{E} - \bar{A}. \tag{11}$$

表达式(11)为集成学习的误差-分歧分解结果[15]，其中 $\bar{E}$ 表示基础学习器的泛化误差之和，$\bar{A}$ 表示基础学习器和集成分类器的差异性之和，也可以看作是个体分类器的多样性。从上述推导可以分析出集成学习的 3 个优化方向：

(1)增加个体分类器的多样性，即产生更多种类的、互相独立的基础分类器；

(2)优化个体分类器的泛化能力；

(3)优化个体分类器的结合策略。

1.3.2　倾向性异质装袋算法

本文选择从优化个体分类器的结合策略、增加个体分类器的多样性两个方面来集成学习算法的优化，提出倾向性异质装袋算法(Propensity Heterogeneous Bagging，PHB)。

在结合策略优化方面，黎竹平通过理论和实验证明了基于 Softmax 函数调整权重的集成学习模型，训练误差要优于不调权的集成学习模型[16]，故本文在此基础上作出改进，使用参数 $\varepsilon \in (0,1]$ 调控集成结果的倾向性，ε 越趋近于 0 则集成学习器越倾向于表现最好的个体学习器，以此来保证集成学习模型的稳定性。计算过程如下：

(1)获得个体分类器的分类准确率 acc 后，将其转换为与 ε -Softmax 函数相关的权重 $\boldsymbol{\omega}$，式(12)：

$$\boldsymbol{\omega}_i = \frac{e^{acc_i/\varepsilon}}{\sum e^{acc_i/\varepsilon}}. \tag{12}$$

(2)根据式(1)计算集成学习器的分类结果。

在增加个体分类器多样性方面，将产生基础学习器的方法分为两大类，一类为相同算法、不同数据集生成基础分类器的同质学习；另一类为使用不同算法、相同数据集生成基础分类器。本文采取多种基础算法和自助重采样(Bootstrap Resamping)相结合的策略来产生基础分类器，在增加个体分类器多样性的同时降低预测准确率的方差。

倾向性异质装袋算法基于多种基础算法和自助重采样(Bootstrap Resamping)相结合的策略来产生基础分类器，在增加个体分类器多样性的同时降低预测准确率的方差，算法的具体过程伪代码如下所示。

倾向性异质装袋算法(Propensity Heterogeneous Bagging，PHB)

输入　训练集 $Data = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$；

基学习算法 $\{f_1, f_2, \cdots, f_n\}$；

倾向程度 ε；

数据重采样次数 T.

过程：

```
for t=1,2,…,T do
    D_t = Bootstrap(Data)
end for
for i = 1,2,…,N do
    for t = 1,2,…,T do
h_i, acc_i = f_i(D_t)
    end for
    ω_i = acc_i/ε
end for
```

输出　$H(x) = \sum Softmax(\omega_i)\, h_i(x)$.

2 实验结果

本文的分析内容为建立预测 S501 钢材表面质量的模型,找出影响钢材质量的关键因素,包括对初始数据进行预处理、特征提取、分析关键工艺参数、评估 PHB 模型的性能。

为分析本文选用的改进集成算法总体的有效性,本文基于相同数据集进行了各类对照实验,实验结果评价的指标为模型运行 100 次后分类准确率的均值(Mean)和方差(Std),以及分类准确率的 95% 置信区间 ($Mean \pm 2 * Std$)。

2.1 特征提取结果

数据集一共有 625 个生产工艺参数、180 条初始样本,在经过缺失值、异常值和归一化等数据预处理后,再使用基于相关性分析和基于随机森林的特征选择方法 RFFS 相结合的方式进行特征选择,确定了影响钢材质量的 39 个特征,其中 23 列确定为强相关特征,16 列为弱相关性特征,具体见表 1。

由于上述特征之间具有高耦合度,故在此基础上再通过主成分分析进行特征融合,最终获得具有 11 列主成分特征的数据集。本文特征提取的流程示意图如图 2 所示。

表 1 最优特征子集

Tab. 1 Optimal feature subset

编号	强相关特征名称	编号	弱相关特征名称
1	出炉温度	24	目标厚度
2	入炉时间	25	测温时间
3	金属离子浓度 g/L	26	大包开浇中包吨位
4	生产日期	27	拉速 1
5	粗轧结束时间	28	空钢包重量
6	精轧结束时间	29	LF 上台时间
7	炉 1 残氧	30	结晶器流量总管压力
8	冶炼时间	31	结晶器流量进水温度
9	AOD 出钢时间	32	二冷水总压
10	出炉平均温度	33	支数
11	软吹开始时间	34	强吹开始时间
12	软吹结束时间	35	强吹结束时间
13	出钢温度	36	中吹开始时间
14	精轧入口实际温度	37	中吹结束时间
15	停浇时间	38	大包开浇时间
16	到台时间	39	R1 支数
17	终轧温度		
18	卷取温度		
19	平均速度		
20	硫酸温度℃		
21	硫酸浓度 g/L		
22	F3 上辊径		
23	F3 下辊径		

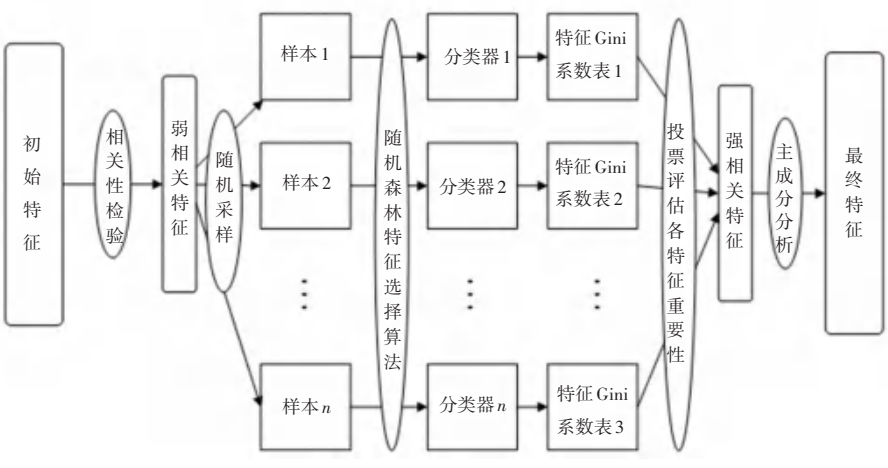

图 2 特征提取流程

Fig. 2 Feature extract process

经过特征提取步骤后获得的 11 列、180 行的主成分数据集中,有 120 条质量合格样本,60 条质量不合格样本。使用分层随机抽样方法抽取 160 个样本为训练集,其余 20 个样本为测试集。

2.2 集成模型的性能验证结果

选择决策树算法、贝叶斯算法、K 近邻算法、逻辑回归算法、BP 神经网络 5 种基础分类算法和 5 次数据重采样,一共生成了 25 个基础分类器,在此基础上进行 ε -Softmax 的权重集成,ε 取值为 0.8。其中 5 种基础分类器预测准确率和 PHB 集成分类器预测准确率对比结果见表 2。

由表 2 可知,基础学习算法具有 80.2~83.8%的分类准确率和 6.7~9.4%的标准差波动,在 95%的置信区间下,某些分类器的准确率有时甚至不如随机猜测,个体分类器之间的分类性能差距较大,故而 PHB 集成学习算法引入 Bagging 策略和倾向性系数 ε 提升预测稳定性,同时使用 Softmax 函数降低个体分类器之间的相关性,提升分类准确率。

表 2　基础分类器和 PHB 算法的泛化性能对比　%

Tab. 2　Comparison of generalization performance between basic classifier and PHB algorithm

	Mean	Std	Mean−2 * std	Mean+2 * std
决策树	82.15	9.40	63.43	100.00
贝叶斯	82.35	7.10	68.18	96.52
K 近邻	80.20	7.60	65.04	95.36
逻辑回归	83.20	8.00	67.29	99.11
BP 神经网络	83.80	6.70	70.45	97.15
PHB 算法	91.40	5.10	81.29	100.00

2.3　结合策略优化分析

为验证基学习器权重结合策略优化的有效性，对 5 种基础分类算法使用了平均权重结合策略和加权结合策略作为对比试验，实验结果如图 3 所示。

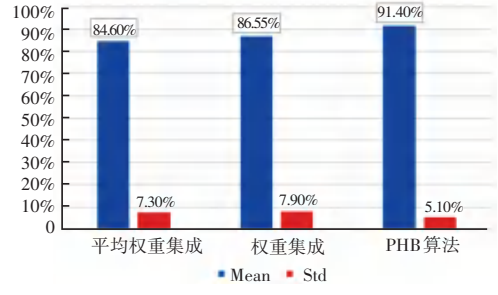

图 3　结合策略优化对比实验

Fig. 3　Comparison of combined strategy optimization

从图 3 可看出，相比于平均权重结合策略 84.6%的平均分类准确率，采用加权结合策略可以使模型平均分类准确率提升至 86.55%，但也会导致标准差的提升。而 PHB 算法不仅使模型提升至 91.4%的分类准确率，同时也降低了模型的方差。

2.4　与常用集成学习算法对比结果

在数据集上选择常规的 3 种 Boosting(AdaBoosting)、Bagging(Random Forest，随机森林)、Stacking(采用 logistic 次级分类器)算法和 PHB 算法进行对比实验，结果见表 3。

表 3　常规集成算法和 PHB 算法的泛化性能对比

Tab. 3　Comparison of generalization performance between conventional ensemble algorithm and PHB algorithm

	Mean	Std	Mean−2 * std	Mean+2 * std
Adaboosting	82.85%	8.30%	66.34%	99.36%
随机森林	80.15%	7.80%	64.50%	95.80%
Stacking	83.85%	8.50%	66.95%	100.00%
PHB 算法	91.40%	5.10%	81.29%	100.00%

由表 3 可见相对于常规的集成学习方法，PHB 集成学习算法不仅提升了约 8%的平均预测准确率，也降低了约 3%的预测准确率标准差，在 95%的置信区间上也可以看出 PHB 模型的稳定性大幅度高于其它算法，适用于实际钢材质量预测场景。

3　结束语

(1)用特征选择算法找出了影响 S501 钢材质量的工艺参数，建立了钢材质量的预测模型，实验表明该模型具有较高精度，分类准确率均值为 91.4%，95%的置信区间为 81.29%~100%，为钢材质量监控模型的建立提供了参考。

(2)使用异质集成以及 Bootstrap 方法增加个体分类器的多样性，降低模型方差、引入倾向性参数和 Softmax 函数优化个体学习器的结合策略两个方面对集成学习器进行了改进，提出基于倾向性的异质装袋算法，并通过实验证明了该优化策略的有效性。

(3)恰当选择倾向性参数 ε 可以提升模型的预测稳定性，如何根据个体学习器的表现来选择对应的倾向性参数 ε 还有待进一步的研究。

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，(2)：75-76.

[2] 吴川平，路同浚，王炎. 钢板表面缺陷的无损检测技术与展望[J]. 无损检测，2002，22(7)：312-317.

[3] 纪英俊，勇晓玥，刘英林，等. 基于随机森林的热轧带钢质量分析与预测方法[J]. 东北大学学报(自然科学版)，2019，40(1)：11-15.

[4] SUI X，LV Z. Prediction of the mechanical properties of hot rolling products by using attribute reduction ELM[J]. The International Journal of Advanced Manufacturing Technology，2016，85(5-8)：1395-1403.

[5] 熊鹰. 带钢表面缺陷检测及识别关键技术研究[D]. 重庆：重庆大学，2016.

[6] 杨威，李维刚，赵云涛，等. 基于随机森林的钢材性能预报与影响因素筛选[J]. 钢铁，2018，53(3)：44-49.

[7] 吴思炜，刘振宇，周晓光，等. 基于大数据的力学性能预测与工艺参数筛选[J]. 钢铁研究学报，2016，28(12)：1-4.

[8] CHIEN C F，WANG W C，CHENG J C. Data mining for yield enhancement in semiconductor manufacturing and an empirical study[J]. Expert Systems with Applications，2007，33(1)：192-198.

[9] 皮骏，马圣，贺嘉诚，等. 遗传算法优化的 SVM 在航空发动机磨损故障诊断中的应用[J]. 润滑与密封，2018，43(10)：89-97.

[10] 江琨，丁学明. 基于集成学习算法的工业产品质量预测[J]. 软件导刊，2019，18(1)：124-127.

[11] 吴思炜，周晓光，曹光明，等. 热轧 C-Mn 钢工业大数据预处理对模型的改进作用[J]. 钢铁，2016，51(5)：88-94，100.

[12] ZHOU Z H. Ensemble learning[J]. Encyclopedia of biometrics，2015：411-416.

[13] ZHOU Z H，WU J，TANG W. Ensembling neural networks：many could be better than all[J]. Artificial intelligence，2002，137(1-2)：239-263.

[14] R. Polikar，Ensemble learning[J]. Ensemble Machine Learning，Springer.2012，1-34.

[15] KROGH A，VEDELSBY J. Neural network ensembles，cross validation，and active learning[C]//Advances in neural information processing systems. 1995：231-238.

[16] 黎竹平. 基于集成学习的特征选择算法的设计与实现[D]. 哈尔滨：哈尔滨工业大学，2018.