文章编号：2095-2163(2021)04-0170-04 中图分类号：TP399 文献标志码：A

基于机器学习的铝及其合金晶粒细化研究

温丽涛，陈 勇
（南华大学 机械工程学院，湖南 衡阳 421000）

摘 要：机器学习作为人工智能的核心，被广泛用于各行各业。而材料基因工程大数据技术的出现，使机器学习成为预测材料性能的新方法。本文利用集成算法 Xgboost、随机森林(RF)以及 AdaBoost 3 种机器学习模型，建立了铝及其合金的晶粒尺寸与细化剂成分、含量的关系，预测铝及其合金的晶粒尺寸，并对该模型进行了交叉验证。结果表明：随机森林模型表现最佳，其测试集的均方根误差(Root Mean Square Error，RMSE)为 7.04，决定系数 R^2 为 0.79。

关键词：人工智能；机器学习；大数据技术；晶粒细化

Research on grain refinement of aluminum and its alloys based on machine learning

WEN Litao，CHEN Yong
（College of Mechanical Engineering，University of South China，Hengyang Hunan 421000，China）

【Abstract】 With the rapid development of artificial intelligence，how to use computer technology to achieve intelligence has become a subject of concern. As the core of artificial intelligence，machine learning is widely used in all walks of life. With the emergence of big data technology in material genetic engineering，machine learning has become a new method to predict material properties. In this paper，three machine learning models of integrated algorithm Xgboost，random forest（RF）and Adaboost were used to establish the relationship between the grain size of aluminum and its alloy and the composition and content of refiner. The grain size of aluminum and its alloy was predicted，and the model was cross-validated. The results show that the random forest model has the best performance in this paper，and the Root Mean Square Error（RMSE）of its test set is 7.04，and the determination coefficient R^2 is 0.79.

【Key words】 Artificial intelligence；Machine Learning；Big data technology；Grain size

0 引 言

大数据与人工智能的结合被称为“科学的第四范式”，机器学习被誉为材料研发的新方法。数据可以从实验、模拟计算、各大材料数据库等获取，再利用机器学习进行数据挖掘，对材料进行研究[1]。机器学习的出现为广大科研人员的科学研究给予了极大的便利，可以加速材料的研究，节约研发成本。目前，已有许多研究人员使用机器学习模型并获得成果。如：胡建军等[2]利用多种机器学习模型，对材料弹性性能进行了归纳和预测；Zhang 等[3]利用支持向量机回归算法，从一系列合金元素中找到恰当合金元素，显著提高铜合金的极限抗拉强度和电导率；Shen 等[4]利用基于不同机器学习算法，结合物理冶金预测材料的强度，成功设计出了超高强度不锈钢。如此等等，都是利用机器学习来对目前产生的大量材料数据进行数据挖掘，利用已有数据对材料性能进行预测，不仅能充分利用材料数据，而且能对实验研究进行系统性的指导。

晶粒细化是提高铝合金性能的重要手段，所以细化剂在工业界被广泛应用[5]。由于晶粒尺寸能很好的表征晶粒细化的结果，因此本文利用机器学习，实现晶粒尺寸的预测。机器学习的晶粒尺寸性能预测模型工作流程如图 1 所示。

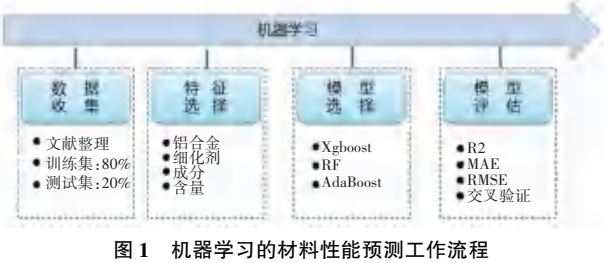

图 1 机器学习的材料性能预测工作流程

Fig. 1 Workflow of performance prediction model via machine learning

本研究首先从文献中收集筛选出需要的数据样

作者简介：温丽涛(1995-)，男，硕士研究生，主要研究方向：机器学习与材料界面研究；陈 勇(1981-)，男，博士，副教授，硕士生导师，主要研究方向：增材制造。

通讯作者：陈 勇 Email：chenyongjsnt@163.com

收稿日期：2021-01-29

本、选择特征、进行数据预处理；然后分别采用 Xgboost[6]、RF[7]以及 AdaBoost[8]3 种机器学习模型预测晶粒尺寸；最后采用五折交叉验证的方式进行不同模型的验证、评估指标、得到最佳模型，从而实现铝及其合金晶粒尺寸的预测。

1　机器学习算法

1.1　Xgboost

Xgboost 是一种集成算法，自提出后便被广泛用于数学界和工业界。其基本思想，是把成百上千个准确率较低的、且合理生成的每颗树，组合成一个准确率较高的模型。该算法有诸多优点：有正则化项防止过拟合、可以加快训练速度等等。

XGBoost 的目标函数如下：

$$Obj = \sum_{i=1}^{n} l(y_i - \hat{y}_i) + \sum_{k=1}^{n} \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2. \quad (1)$$

式中：$\hat{y}_i$ 为真实值；y_j 为预测值；T 表示叶子结点的个数；w 表示叶子节点的分数；γ、λ 为校正常数。

1.2　随机森林

RF 是由 Leo Breiman 于 2001 年提出，由决策树组合成的集成算法。即由很多决策树组成的森林，且每棵决策树之间无关联。RF 是指在变量和数据的使用上进行随机化，产生很多决策树，再将之汇总的结果。RF 的构造过程如下：

(1)用 N 表示训练集的个数，M 表示特征数目；

(2) 输入特征数目 m，用于确定决策树节点的结果，m 应远小于 M；

(3) 从 N 个训练集中以有放回抽样的方式，取样 N 次，形成一个训练集，并用未抽到的样本作预测，评估其误差；

(4) 对于每个节点，随机选择 m 个特征，计算其最佳的分裂方式；

(5) 每棵树都不会剪枝，这有可能在建完一棵正常树状分类器后会被采用；

(6)按照步骤(3)～(5)建立大量的决策树，从而构成 RF。

1.3　AdaBoost

Adaboost 算法是一种自适应的集成算法，1995 年由 Yoav Freund 和 Robert Schapire 提出。它的自适应在于：如果前一个弱分类器将样本分错，那么样本对应的权值会得到加强，即权值更大，权值更新后的样本用来训练下一个新的弱分类器。每次训练时，都是用总样本来训练新的弱分类器，产生新权值，如此反复迭代直到达到预定的评估指标或达到最大迭代次数。简要过程如下：

(1)初始化训练集的权值分布。如果有 n 个样本，则每一个训练的样本点最开始时都被赋予相同的权重为 $1/n$；

(2)弱分类器训练。具体训练过程中，若样本准确分类，则权值降低，否则，权值增大，即弱分类器得到更高话语权。权值更新后的数据集被用于训练下个分类器，如此反复迭代；

(3)集成算法。将多个弱分类器组合成强分类器，在每个弱分类器的训练过程结束后，误差率高的弱分类器在最终分类器中占的比例较小，反之则较大。

2　机器学习预测晶粒尺寸

2.1　数据准备

本文通过对已有文献中的铝及其合金的相关数据的查找，收集样本数据[9]。选择的特征变量包括铝及其合金类型、细化剂类型、细化剂成分和细化剂比例，其中晶粒尺寸的值作为目标变量，即预测的目的变量。

2.2　软件选择

本文使用 Python3.7 进行数据处理和算法模型计算。Python 作为一款开源软件，因其代码的可读性高，且拥有丰富强大的扩展包。其中 ，第三方 Scikit-learn[10]库，集成了各种数据处理方法及大量算法模型，可高效便捷地进行数据处理和建立机器学习模型。文中使用的 One-Hot 编码、数据标准化、RF 以及 AdaBoost 算法都选自 Scikit-learn 包。Xgboost 算法来自第三方扩展包。

2.3　数据处理

预测晶粒尺寸时，需进行特征数据的预处理。因铝及其合金与细化剂种类为非数值类型的数据，无法参与计算，在建立模型之前，需将其进行独热编码，也就是 One-Hot 编码。

由于不同特征之间的数据量级相差较大，所以还需要将特征数据进行标准化处理。处理后的数据不仅可以消除数据量级不一致对机器学习带来的影响，并使数据仍保持了原始分布。即对每个特征进行如下变换：

$$x_i = \frac{x_i - \bar{x}}{\sigma}. \quad (2)$$

式中，x_i 为输入的数据，$\bar{x}$ 和s分别为特征的均值和标准差。值得注意的是，该处理虽然使原本数据失去本来的意义，但利于模型的建立。

2.4　模型评估

本文使用平均绝对误差 MAE、均方根误差

(Root Mean Squared Error, RMSE)和决定系数 R^2 值被用来作为泛化性能评估。公式如下:

$$E_{MAE} = \frac{1}{n}\sum_{j=1}^{n} | y_j - \hat{y}_i | , \quad (3)$$

$$E_{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (\hat{y}_i - y_j)^2} , \quad (4)$$

$$R^2 = 1 - \frac{\sum_{j=0}^{n-1} (y_j - \hat{y}_i)^2}{\sum_{j=0}^{n-1} (y_j - \bar{y}_j)^2} . \quad (5)$$

式中, n 为样本数量; $\hat{y}_i$ 为真实值; y_j 为预测值。R^2 为决定系数的最大取值,设定为 1。取值越接近 1,表明拟合程度越好。

3 结果分析

不同模型的测试集评估结果见表 1。其中,3 种机器学习模型的 R^2 值相差不大,表现最好的是 Xgboost 模型, R^2 为 0.929 9(即可以解释 92.99 的方差), *MAE* 和 *RMSE* 分别为 6.220 9、38.699 0。AdaBoost 表现结果次之,RF 相对最差。由此可见,对于同一数据集,不同的机器学习模型表现效果不一致。因此,合理选用机器模型对预测结果十分重要。

表 1 不同机器学习模型下测试样本的预测结果比较

Tab. 1 Comparison of test sample results under different machine learning models

模型	*MAE*	*RMSE*	R^2
Xgboost	38.699 0	6.220 9	0.929 9
RF	53.102 8	7.287 2	0.887 8
AdaBoost	58.729 7	7.663 5	0.904 9

为了充分使用数据集,实验采用五折交叉验方法,保证每个数据都能作为训练集或测试集出现。Xgboost、RF、AdaBoost 算法的 R^2 交叉验证结果见表 2。可以看出,3种算法在交叉验证中的R^2均存在波动,说明在选取不同的训练集与测试集时,会导致预测结果有较大差异。其中 RF 算法在交叉验证的 5 折交叉验证中,其 R^2 预测结果在 0.50-0.97 之间波动,变化起伏相对其它两个较小,说明该算法在应对数据集时稳定效果最好,且预测结果较好。Xgboost、AdaBoost 的 R^2 交叉验证结果相对 RF 来说波动幅度更大,说明不同算法应对不同属性数据的稳定效果不同。图 2 为不同机器学习模型的 R^2 与 RMSE 交叉验证均值结果,可以看出误差都比较大,预测结果不稳定,可能是数据量不足所致。其中 RF 的 R^2 值为 0.79,RMSE 的值为 7.04,说明在 3 种机器学习模型中预测结果最佳。

表 2 不同机器学习的 R^2 交叉验证结果

Tab. 2 Cross validation of R^2 results under different machine learning

	五折交叉验证				
Xgboost	0.977 6	0.768 5	0.237 1	0.538 9	0.959 7
RF	0.969 4	0.796 4	0.509 6	0.747 6	0.943 8
AdaBoost	0.145 3	0.686 6	0.680 5	0.691 4	0.937 8

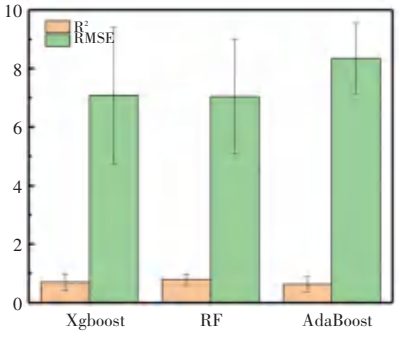

图 2 不同机器学习模型下 R^2 与 RMSE 交叉验证均值结果

Fig. 2 Cross validation mean results of R^2 and MAE under different machine learning model

为了更直观地说明 3 种不同机器学习模型的优劣,比较了不同算法下的预测结果,如图 3 所示。可以看出,3 种机器学习模型的预测结果大多在 $y = x$ 之上或附近,即预测结果相对真实值相吻合,其中 RF 的预测结果最佳。

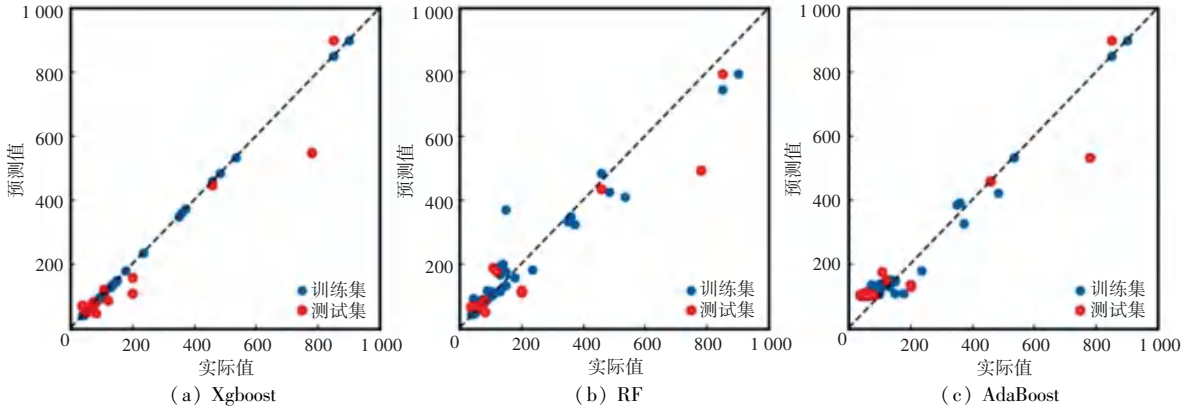

图 3 不同机器学习下的预测结果

Fig. 3 Prediction results under different machine learning

(下转第 175 页)