

基于 EXAFS 谱图的机器学习方法预测配位数

曾海滔 胡龙飞 姚 涛

(中国科学技术大学 国家同步辐射实验室 合肥 230029)

摘要 X射线吸收精细结构(X-ray Absorption Fine Structure, XAFS)是一种重要的结构分析技术,广泛应用于研究非晶态材料和无序体系的氧化态、配位环境及邻近原子特性。然而,由于 XAFS 谱图的复杂性,其解析依赖于经验丰富的科研人员,且易出现不准确性。本研究利用机器学习方法分析 XAFS 数据,特别是扩展 X 射线吸收精细结构(Extended X-ray Absorption Fine Structure, EXAFS)谱图,以预测吸收元素的配位数。从 Materials Project 数据库中收集了第四周期过渡金属元素的 13 374 条有效 EXAFS 数据,涵盖多种配位环境。利用这些数据训练神经网络、袋装法决策树模型和随机森林模型,应用这些模型预测吸收原子的配位数。结果表明:三种模型的平均预测准确率均达到约 70%,特征重要性分析显示 $R < 0.3$ nm 的数据点对预测至关重要。本研究通过提升模型的泛化能力和可解释性,为 XAFS 数据分析提供了更高效、可靠的方法。

关键词 X射线吸收精细结构, 配位数, 袋装法, 随机森林

中图分类号 TL99, O722

DOI: 10.3724/j.0253-3219.2026.hjs.49.250117

CSTR: 32193.14.hjs.CN31-1342/TL.2026.49.250117

Machine learning approach for predicting coordination numbers from EXAFS spectra

ZENG Haitao HU Longfei YAO Tao

(National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei 230029, China)

Abstract [Background] X-ray Absorption Fine Structure (XAFS) is a vital technique for structural analysis, but the complexity of XAFS spectra often requires interpretation by experienced researchers, which can still lead to inaccuracies. [Purpose] This study aims to develop machine learning models for predicting coordination numbers directly from Extended X-ray Absorption Fine Structure (EXAFS) spectra of fourth-period transition metal elements. [Methods] Firstly, a dataset of 13 374 valid EXAFS spectra was collected from the Materials Project database, covering diverse coordination environments of fourth-period transition metals. Secondly, the EXAFS data were transformed from energy space to R-space using Fourier transformation, with intensity values extracted at 0.003 nm intervals as features. Thirdly, three machine learning models, i. e., neural networks, bagging decision trees, and random forests, were trained and evaluated using five-fold cross-validation to predict the coordination numbers of absorbing atoms. Finally, feature importance analysis based on the Gini index was performed to identify key spectral regions contributing to predictions. [Results] Prediction results show that three models achieve an average prediction accuracy of approximately 70%, with the neural network reaching up to 81.74% for vanadium. Feature importance analysis reveals that data points at $R < 0.3$ nm are critical for coordination number predictions, accounting for the

国家自然科学基金(No.12025505)资助

第一作者: 曾海滔, 男, 1997 年出生, 2020 年毕业于中国科学技术大学, 现为硕士研究生, 研究领域为核科学与技术

通信作者: 姚涛, E-mail: yaot@ustc.edu.cn

收稿日期: 2025-04-03, 修回日期: 2025-10-30

Supported by National Natural Science Foundation of China (No.12025505)

First author: ZENG Haitao, male, born in 1997, graduated from University of Science and Technology of China in 2020, master student, focusing on nuclear science and technology

Corresponding author: YAO Tao, E-mail: yaot@ustc.edu.cn

Received date: 2025-04-03, revised date: 2025-10-30

majority of predictive power, which is consistent with the dominance of short-range atomic interactions in EXAFS theory. **[Conclusions]** Results of this study demonstrate that machine learning models, particularly random forests with Gini-based feature analysis, provide interpretable and reliable tools for EXAFS data analysis. The achieved ~70% average accuracy across diverse transition metal systems indicates strong generalizability, while the identification of critical spectral features ($R < 0.3$ nm) validates the physical basis of model predictions and enhances the efficiency of automated structural characterization.

Key words X-ray absorption fine structure, Coordination number, Bagging, Random forest

X射线吸收精细结构(X-ray Absorption Fine Structure, XAFS)可分为两个主要区域:X射线近边结构(X-ray Absorption Near Edge Structure, XANES)以及扩展边X射线吸收精细结构(Extended X-ray Absorption Fine Structure, EXAFS)。XAFS对吸收原子的氧化态、配位环境以及邻近原子的种类与距离具有高度敏感性。因此,其在物理学、化学、材料科学以及生物学等多个学科领域中得到广泛应用。然而,由于XAFS谱图的复杂性,其解析通常依赖于具有丰富经验的科研人员。即便如此,专业人员在分析过程中仍可能得出不准确的结论^[1]。因此,开发更为高效和可靠的分析方法以提升XAFS数据的解析精度,已成为当前研究的重要方向。

机器学习作为近年来最具代表性的数据驱动研究方法,已成为科学研究领域的重要热点。其中,神经网络以其在预测任务中表现出的高准确率而备受关注。当前神经网络已经被广泛应用于扩展边光谱解析、近边结构特征提取等课题。例如,Tetef等^[2]证明,无监督方法如 t -分布随机邻域嵌入(t -distributed Stochastic Neighbor Embedding, t -SNE)和变分自编码器(Variational Autoencoder, VAE)可基于XANES和价带-内层X射线发射光谱(Valence-to-Core X-ray Emission Spectroscopy, VtC-XES)光谱有效分类硫有机化合物,揭示氧化态和芳香性等详细化学特性;Timoshenko等^[3]开发了一种基于神经网络的方法,直接从EXAFS光谱中提取径向分布函数(Radial Distribution Function, RDF),无需依赖先验结构假设。该方法成功应用于铁(Fe)的高温体心立方(body-centered cubic, bcc)到面心立方(face-centered cubic, fcc)相变分析,并推广至钴(Co)、镍(Ni)等体系。

然而,现有的相关工作通常局限于特定材料或材料类别在不同参数(如温度变化)下的简单材料体系分析,存在模型泛化的困难。此外,由于神经网络的复杂性以及其“黑箱”特性,研究人员往往难以解析模型生成参数的具体依据,同时无法直观理解其决策过程。这种可解释性缺失限制了神经网络在需

要明确特征重要性分析的研究场景中的应用^[4]。为了提高机器学习模型的可解释性,研究者开发了多种技术方法,包括岭回归(Ridge Regression, RR)^[5]、最小绝对收缩和选择算子(Least Absolute Shrinkage and Selection Operator, LASSO)回归^[6]、SISSO(Sure Independence Screening and Sparsifying Operator)方法^[7]以及决策树(Decision Tree, DT)^[8]等。其中,决策树因其模型结构简单且具有较高的可解释性而受到广泛关注。此外,基于集成学习(Ensemble Learning, EL)^[9],研究者进一步发展出了袋装法(Bagging)^[10]、随机森林(Random Forest, RF)^[11]、提升树(Boosting)^[12]等方法。决策树模型在XAFS谱学研究领域已展现出了其潜力。例如Torrisi等^[13]利用随机森林机器学习模型分析XANES谱,预测过渡金属氧化物中吸收原子的配位数、最近邻距离和Bader电荷。研究通过多尺度多项式特征化方法提高了模型的可解释性,揭示了光谱中与不同性质相关的关键区域。

针对机器学习在XAFS解析中的两大核心制约问题,即模型泛化困难和可解释性缺失,本研究利用Materials Project数据库构建了一个覆盖第四周期多种过渡金属元素、跨越不同材料类别和多样化配位环境的大规模EXAFS数据集,基于此数据集开发的神经网络和随机森林模型,在较高精度的配位数预测上展现出了泛化能力。此外,随机森林模型的可解释性分析成功识别了决定配位数的谱学特征区域,即傅里叶变换后低 R 区域。本研究提供了一种可扩展、强泛化能力的XAFS数据分析工具,为广泛材料体系的高通量、自动化XAFS结构分析奠定了坚实基础,具有重要的方法学价值和实用前景。

本课题的流程如图1所示。从Materials Project中收集了共计15 498条第四周期过渡金属元素的XAFS数据,经过筛选,保留了配位数为整数的13 374条数据。随后,将这些XAFS数据从 E 空间转换到 R 空间,并分别将其导入全连接神经网络以及决策树的袋装法与随机森林集成学习模型中。对于每种元素,我们分别训练模型并进行评估。

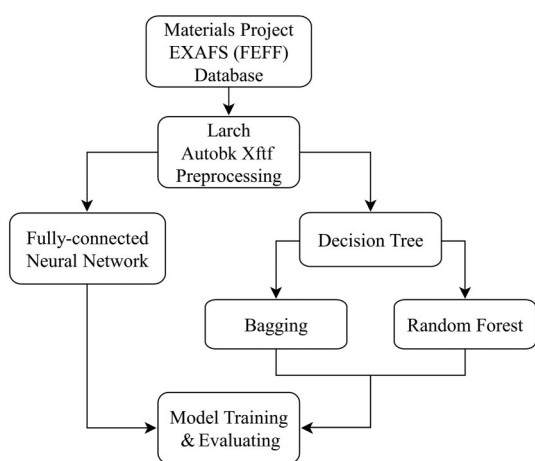


图1 本研究的流程图
Fig.1 Flowchart of this study

1 背景

1.1 X射线吸收精细结构谱(XAFS)及其扩展部分(EXAFS)

X射线吸收精细结构(XAFS)现象于1920年由Fricke与Hertz首次在实验中发现并予以科学记载。对XAFS研究的突破性进展源于Sayers、Stern及Lytle^[14]三位学者提出的创新方法——通过傅里叶变换技术对扩展X射线吸收精细结构(EXAFS)振荡信号的波矢空间进行有效解析。这一里程碑式的工作标志着XAFS技术正式进入定量分析领域,并开启了其在材料表征方面的系统化发展进程^[15]。

XAFS技术能够提供关于吸收原子周围局部化学环境的详细结构信息。近年来,基于XAFS的原子级局部性质分析已成为材料科学和催化研究中的一种重要手段。具体而言,XAFS技术被应用于金属团簇的配位环境研究、铁电材料中局部结构畸变的探讨,以及单原子催化剂的催化机理分析等多个课题。杨旺等^[16]利用EXAFS技术研究针铁矿和赤铁矿对As(V)的吸附,揭示了As与铁氧化物形成的络合物的分子结构。陈雨等^[17]根据同步辐射XAFS荧光测量实验中的需求,设计了一款荧光气体电离室探测器,超过了进口设备的性能指标。

XAFS是通过测量材料对X射线的吸收系数随入射光子能量的变化而获得的光谱。XAFS谱通常可分为两个特征区域:XANES和EXAFS。EXAFS的理论表达式可表示为:

$$\chi(k) = S_0^2 \sum_j \frac{N_j f_j(k) e^{-2k^2 \sigma_j^2} e^{-2R_j/\lambda(k)}}{k R_j^2} \sin[2k R_j + \delta_j(k)] \quad (1)$$

式中: k 表示光电子波矢; N_j 表示第 j 配位壳层的配位数; R_j 表示吸收原子到第 j 配位壳层邻近原子的平均距离; σ_j^2 表征第 j 配位壳层的无序度; $f_j(k)$ 表示邻近原子对光电子的散射能力; $\delta_j(k)$ 表示相位位移; $\lambda(k)$ 表示光电子传播的平均自由程; S_0^2 表示振幅衰减因子。

上述理论表达式表明,通过傅里叶变换可以将EXAFS振荡函数从 k 空间转换到 R 空间,从而获得关于吸收原子周围局域结构的径向分布信息。然而,由于EXAFS公式中同时存在多个相互耦合的可变参数,以及可能存在多条散射路径的叠加效应,仅依靠EXAFS谱图进行精确的结构解析面临严重的不适定问题。因此,在实际的EXAFS数据分析中,研究者通常采用“指纹比对”的方法,即通过将实验数据与已知结构的参考谱图或理论模拟结果进行对比,从而获得材料局域结构的相对模型。

尽管直接从XAFS谱图反演所有结构参数存在固有困难,但基于已知结构模型计算理论XAFS谱图已成为一种可靠的研究方法。在XAFS研究领域,FEFF^[18]和FDMNES^[19]是两种被广泛采用的计算软件包。这些软件均基于第一性原理(ab initio)计算方法,通过求解薛定谔方程来精确模拟XAFS谱图特征。其中,FEFF采用实空间多重散射理论(Real-Space Multiple Scattering, RSMS),而FDMNES则基于有限差分方法(Finite Difference Method, FDM),为实验数据的解析提供可靠的理论参考。

1.2 Larch软件包

Larch是一个专门用于XAFS及其相关光谱数据分析的综合工具包^[20]。与基于FORTRAN语言开发的IFEFFIT软件包相比,Larch采用Python语言进行重构,这显著提升了大规模数据处理能力和数据可视化功能。在本研究中,我们利用Larch工具包中的核心函数进行数据处理:其中,autobk函数用于实现背景扣除和归一化处理,将EXAFS谱图从 E 空间转换到 k 空间;而xftf函数则用于执行傅里叶变换(Fourier Transform, FT),将EXAFS谱图从 k 空间转换到 R 空间。

1.3 配位数

配位数是描述材料中原子局部配位环境的关键结构参数,对于研究材料的性能具有至关重要的意义。在传统研究中,配位数通常通过直观判断来确定材料中单个原子的配位情况。然而,随着高通量材料计算技术的发展,产生了大规模数据集,因此开发可靠且自动化的配位数计算算法已成为研究的迫

切需求。

目前,材料科学领域提出了多种配位数计算方法^[21]。其中,较为简单的算法基于原子之间的距离以及相关容忍度的经验参数,如 BrunnerNN^[22]和 MinimumOKeeffeNN^[23]。这些方法通过比较模型中原子之间的实际距离及其容忍值来确定配位数。尽管这些算法较为简便且直观,但它们对原子微小扰动和经验容忍参数的变化较为敏感,原子的微小变化可能导致配位数的不同分配。

O'Keeffe^[24]提出的 VoronoiNN 算法提供了另一种解决方案。该算法基于几何原理,采用 Voronoi 分解^[25]将中心原子周围的环境视为一个多面体,并进一步根据空间角度进行加权处理,从而得出近邻原子的数量。与此类似,本研究使用了 CrystalNN^[26]算法。CrystalNN 方法同样基于 Voronoi 分解,并因其较高的计算精度而被广泛认可。该方法通过 Voronoi 分解计算可能的配位环境的概率,最终选取概率最高的结果作为该原子的配位环境。在本文中,我们借助 pymatgen 材料计算软件包^[27]中的 CrystalNN 算法,系统地计算了各材料体系中吸附原子的配位数。

1.4 神经网络

作为机器学习的重要分支,神经网络(深度学习)技术因其在回归和分类任务中展现出的卓越性能,已成为当前科学研究和技术应用的核心领域之一。神经网络模型通过大量可调参数实现特征提取和模式识别,这些参数在训练过程中通过反向传播算法自动优化。研究者主要通过调节超参数(包括批量大小、学习率和网络层结构等)来优化模型性能。

单层隐藏层前馈神经网络模型可形式化定义为式(2):

$$f(X) = \beta_0 + \sum_{k=1}^K \beta_k g\left(w_{k0} + \sum_{j=1}^p w_{kj} X_j\right) \quad (2)$$

式中: $X = (X_1, X_2, \dots, X_p) \in R^p$ 表示 p 维输入特征向量。模型参数包括: $w_{kj} \in R$ 表征输入层第 j 个特征到隐藏层第 k 个神经元的连接权重, $\beta_k \in R$ 表示隐藏层第 k 个神经元至输出层的权重系数,而 $w_{k0} \in R$ 和 $\beta_0 \in R$ 分别为对应线性变换的偏置项。

尽管该模型架构在形式上呈现为线性叠加结构,但为有效逼近现实世界中的非线性函数关系,需引入非线性激活函数 $g(\cdot)$ 。在机器学习领域,Sigmoid 函数与 ReLU (Rectified Linear Unit)^[28] 函数是两类经典激活函数。Sigmoid 函数具有 S 型饱和特性,其输出值域为 $(0, 1)$,这一特性使其在早期神

经网络研究中获得广泛应用。然而,该函数在输入值趋于正负无穷时会出现梯度消失现象,即导数趋近于零,导致反向传播过程中参数更新效率显著降低。相较而言,ReLU 函数定义为 $g(x) = \max(0, x)$,其分段线性特性保证在正区间内导数为恒定值,从而有效缓解梯度消失问题。此外,ReLU 函数的计算复杂度显著低于 Sigmoid 函数,这使得其在深层神经网络训练中展现出更优的计算效率与收敛特性。

1.5 决策树

尽管神经网络在预测性能方面具有显著优势,但其固有的“黑箱”特性导致模型可解释性不足。具体而言,神经网络中大量神经元参数的设置机制难以被清晰解读,这一局限性促使研究者寻求其他可解释性更强的辅助分析工具。在本研究中,我们采用决策树作为分析工具,其优势在于模型结构直观且易于解释。通过决策树,研究者可以清晰地识别数据特征与结果之间的关联关系。然而,传统决策树方法在预测准确性方面存在明显不足。为克服这一局限,研究者开发了多种集成学习方法,包括袋装法、随机森林和提升树等。这些集成方法虽然显著提升了预测性能,但牺牲了单一决策树的直观解释性。尽管如此,通过分析基尼指数(Gini Index),我们仍可有效评估各特征在分类决策中的重要性。

决策树的建模机制遵循递归式特征空间划分策略。该算法通过预设的树形结构超参数(如最大深度),采用监督学习方式从训练数据集中提取最优分裂准则,从而将 p 维特征空间(假设输入样本为 p 维特征向量)分解为若干互斥的几何子空间。理论研究表明,该模型对具有显式超平面划分特征的数据集展现出优异的分类性能,但在处理非线性可分数据或具有复杂决策边界的分类问题时,其分类效能将呈现显著下降趋势。然而,实证研究表明,在处理表格型数据时,基于决策树架构衍生的集成学习算法相较于神经网络模型展现出更优越的预测精度。以 XGBoost 算法为例,该优化型梯度提升框架通过迭代式决策树集成与正则化策略,在多个基准测试中取得了显著的性能突破^[29]。

1.6 材料数据库

在材料科学领域的机器学习研究中,构建稳健的预测模型高度依赖于大规模、高质量的数据集支撑^[30]。当前主流的材料信息学平台包括 Open Catalyst^[31]、Materials Project^[32],以及开放量子材料数据库(Open Quantum Materials Database, OQMD)^[33]等。本研究选择 Materials Project 作为数据源,主要基于以下考量:该数据库不仅整合了超过

16 万种材料的结构信息,还提供了基于 FEFF 理论计算的 XAFS 光谱数据,这对于建立材料构效关系具有重要价值。此外,其开放的 Python 应用程序接口 (Application Programming Interface, API) 显著提升了数据获取效率,并能与高通量计算 workflow 无缝衔接。通过 pymatgen 材料分析工具包进行数据获取,所有处理后的结构化数据均采用 MongoDB 文档数据库进行存储,以实现高效检索。

根据 Materials Project 的官方文档引用的文献,FEFF9 中的输入字段进行了收敛检查以及优化测试:在收敛检查中,研究人员通过将 r_{fms} 值从 0.2 nm 变化到 0.8 nm,间隔 0.1 nm 改变 FEFF 的自洽势自动计算。自洽势由自洽场 (Self-Consistent Field, SCF) 卡控制。同时,将 r_{fms} 值从 0.3 nm 变化到 1.1 nm,间隔 0.1 nm 来确定全多重散射 (Full Multiple-Scattering, FMS) 卡的参数。

2 数据集与预处理

本研究的数据来源于 Materials Project 数据库,我们系统检索了包含第四周期过渡金属的材料,并从中提取了 EXAFS 计算数据。所有 EXAFS 数据均通过 FEFF 计算软件包生成。为确保数据质量,我们排除了同一材料中存在不同配位数吸收原子的样本,因为这些数据可能会对配位数分类产生干扰。

在本研究中,我们使用 Materials Project 数据库中理论计算的 EXAFS 数据而不是加入噪音计算的 EXAFS 数据。根据 Ghigna 等^[34]以及 Marcella 等^[35]的研究,给计算数据加上噪音所造成的影响小于系统因素(如背景减去等数据处理)的影响。因此可以直接使用计算出的 EXAFS 数据进行机器学习。

数据库中包含多种复杂体系的材料,比如氧化物、硫化物以及合金等。这能增强模型的泛化能力,

但会对分类造成挑战。同时,过于复杂的分类会稀释单一类别的预测精度。因此,本研究侧重于只对吸收原子进行分类。

针对具有周期性的 EXAFS 数据,采用傅里叶变换进行预处理。对于大规模数据处理,使用 Larch 工具包结合 Python 编程语言实现自动化处理。具体而言,利用 Larch 的 autobk 函数将原始 E 空间数据转换为 k 空间数据,随后通过 xftf 函数将 k^2 加权的 k 空间数据傅里叶变换为 R 空间数据。

在 R 空间数据处理方面,我们以 0.003 nm 为间隔提取强度值作为特征。此外,采用 Python 的 Peak 数据包计算每个特征峰对应的 R 坐标和强度值,并将这些参数作为附加特征纳入数据集。

为进行对比,我们分别使用 k 空间数据和小波变换后的数据训练神经网络模型。小波变换采用 Larch 工具包中的 cauchy_wavelet 函数,参数设置为 $k_{\text{weight}}=0$, $r_{\text{max_out}}=6$,对 $k_{\text{min}}=3$ 至 $k_{\text{max}}=14$ 的 k 空间数据进行变换。

3 模型构建与训练

为完成配位数预测任务,本研究构建了基于全连接神经网络的预测模型。如图 2 所示,神经网络的输入层包含 326 个特征,这些特征对应于 EXAFS 的 R 空间谱图在间距 0.003~0.003 1 nm 下的强度值。该网络采用双隐藏层结构,每层包含 512 个神经元。各隐藏层后依次配置 ReLU 激活函数层和 Dropout 正则化层:激活函数采用 ReLU 以增强模型的非线性表达能力,同时引入 Dropout 正则化机制以缓解过拟合问题。输出层的参数 k 的范围会根据不同元素对应的配位数范围而变化。在模型优化过程中,通过五折交叉验证筛选出验证集性能最优的模型作为最终模型,该方法有效提升了模型的泛化能力。

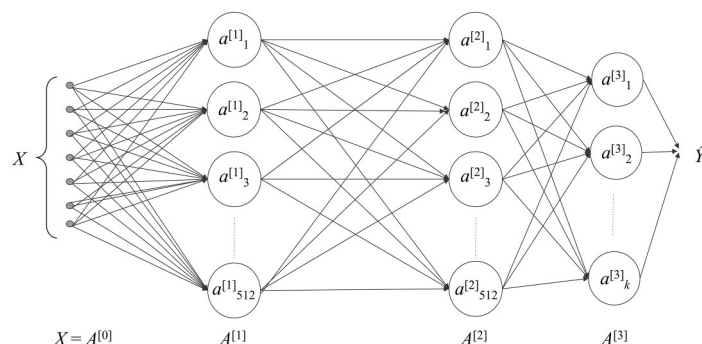


图2 神经网络结构示意图
Fig.2 Structure diagram of the neural network

为提升模型可解释性并验证神经网络的预测结果,本研究同时构建了基于决策树的集成学习模型。

首先采用 Scikit-learn^[36]机器学习框架实现了两种集成模型:基于袋装法的决策树集成和随机森林。袋

装法通过自助采样(Bootstrap Sampling)生成与数据量相等数量的子训练集,每个子集训练独立的决策树模型,最终通过投票机制整合预测结果。随机森林则在袋装法基础上引入特征随机性——每次节点分裂时仅随机选取 $\sqrt{d}$ 个特征(d 为总特征数)进

行最优划分,以此提升模型多样性。两类模型使用基尼指数作为标准。如图3所示,图中的每个方框代表一个分裂节点。方框中的“Freq_r”表示EXAFS谱图在 R 空间中横坐标 r 对应的强度,而“samples”则表示在分裂之前该节点所包含的数据量。

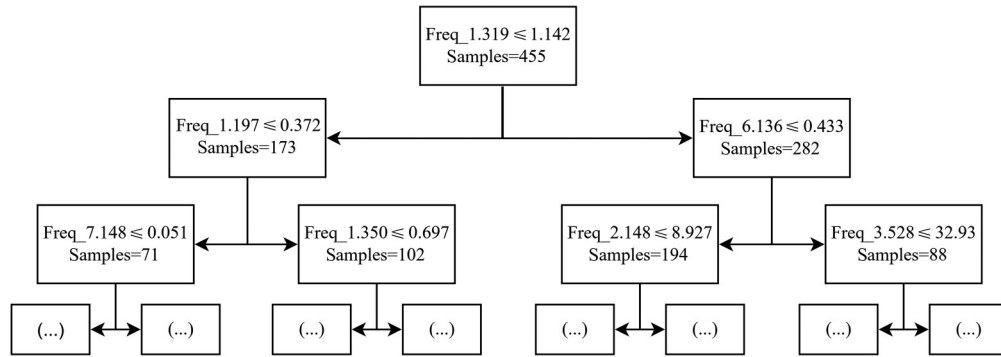


图3 决策树模型示意图
Fig.3 Diagram of decision tree model

4 结果

本研究采用神经网络与决策树算法基于第四周期过渡金属材料的EXAFS谱图进行了系统分析。如图4所示,对于每种金属,模型按从左至右的顺序依次为:神经网络模型、袋装法模型、随机森林模型、结合峰高与峰位置的袋装法模型,以及结合峰高与峰位置的随机森林模型。计算结果表明,两种方法的预测准确率在70%左右,其中神经网络模型表现出略优于决策树模型的预测性能。值得注意的是,两种方法的准确率呈现显著相关性,其数值随元素种类变化呈现同步增减趋势。在集成学习方法的应用中,随机森林模型的预测准确率较袋装模型平均提升1%~3%。此外,本研究还考察了包含峰位置与

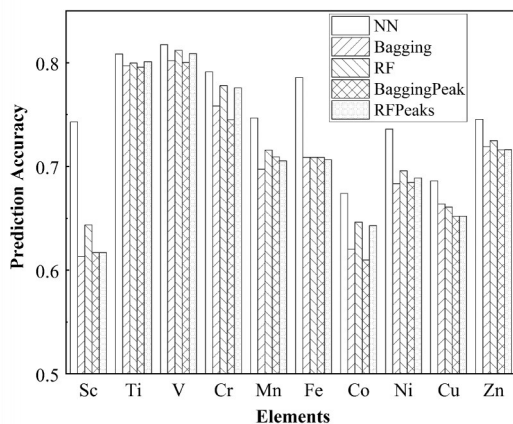


图4 第四周期过渡金属材料配位数预测的准确率
Fig.4 Prediction accuracy of coordination numbers for fourth-period transition metal materials

峰高信息的特征数据集,实验发现决策树模型在该数据集上的预测准确率与原始数据集相比未呈现显著差异。

研究结果表明,在配位数分类问题上,模型的预测准确率呈现出显著的元素依赖性。其中,钒(V)元素在神经网络框架中表现出最优的预测性能,准确率达到81.74%;相比之下Co元素的预测准确率相对较低,仅为67.39%。通过进一步分析,我们发现这一差异可能与元素在数据库中的配位数分布特征有关。具体而言,V元素在大多数材料中保持相对稳定的六配位结构,而Co元素则表现出更为多样化的配位数变化,这可能是导致其预测准确率相对较低的主要原因。

如图5所示,将每个元素的预测率对配位数的信息熵进行作图。散点图采用最小二乘法线性回归

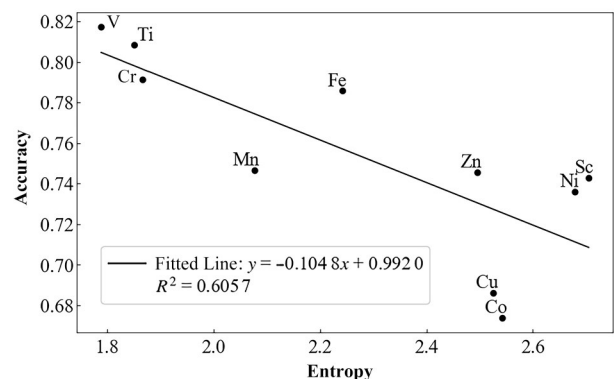


图5 准确率与信息熵的相关性
Fig.5 Correlation between prediction accuracy and information entropy

拟合,揭示了这两个变量之间的关系, R^2 平方值表明了拟合的强度。可以看出,预测准确率与配位数的分布特征呈现出一定的相关性。

同时,本实验对比了基于 k 空间和 q 空间(小波变换)数据的预测准确率。如图6所示,三种空间的预测准确率各有差异,但总体趋势大致相同,即预测率随着元素的变化呈现共同的上升或下降趋势。 R

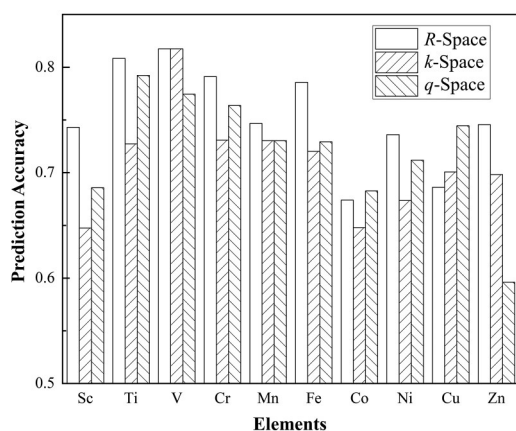


图6 R 空间、 k 空间以及 q 空间对应的预测准确率
Fig.6 Prediction accuracy corresponding to R -space, k -space, and q -space

空间的预测准确率在大多数情况下优于 k 空间和 q 空间,因此我们推荐使用 R 空间进行神经网络学习。

为了展示模型的偏差特性,我们以 Zn 元素作为样例,作出真实配位数和预测配位数对应的混淆矩阵,如图7所示。

本研究进一步基于决策树集成模型的基尼指数,构建了各元素 EXAFS 谱图特征重要性分布图谱。图8展示了 Co 和 Mn 元素 EXAFS 数据在袋装法和随机森林模型中的特征重要度及其与 R 空间的对应关系。由于 R 空间数据集包含 326 个等间距坐标点的强度值,最大单特征重要性值普遍低于 0.1。根据 EXAFS 谱图处理的经验,我们只探讨 $R < 0.6$ nm 的区域。通过对比相对重要性分布发现,特征重要性权重主要集中于 $R < 0.3$ nm 的近程作用区,这与现有 EXAFS 理论中近程原子间相互作用主导谱图特征的经验认知相吻合。

对于 $R > 0.6$ nm 的远距离配位层,目前的模型难以可靠地建立起高配位层的配位数与 R 空间参数的直接对应关系。因为高壳层信号强度较低,导致预测误差显著增大,无法满足准确预测的要求。在实践操作中,我们也很难获取有关高配位层的配位信息。

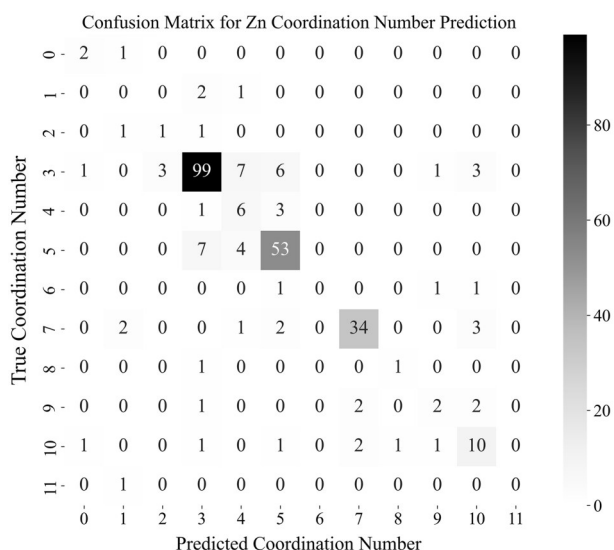


图7 Zn 元素的真实配位数与预测配位数对应的混淆矩阵
Fig.7 Confusion matrix of predicted vs. true coordination numbers for Zn elements

此外,本研究还观察到,袋装法在特征重要性分布图中显著放大了不同特征之间的重要性差异,而随机森林模型中的特征重要性分布则相对均衡。这一现象可能与袋装法模型的过拟合倾向有关。具体而言,袋装法在每次迭代中仅去除一个特征进行优化,导致重要性较高的特征在多次迭代中其重要性被进一步放大。相比之下,本研究中采用的随机森

林模型在每次迭代中同时筛选出 14 个特征进行优化,从而避免了单一特征重要性的过度增强,有效提升了模型的泛化能力。这一发现表明,随机森林模型在处理高维数据时能够更好地平衡特征重要性,减少过拟合风险,从而为 EXAFS 谱图分析提供了更为稳健的解决方案。

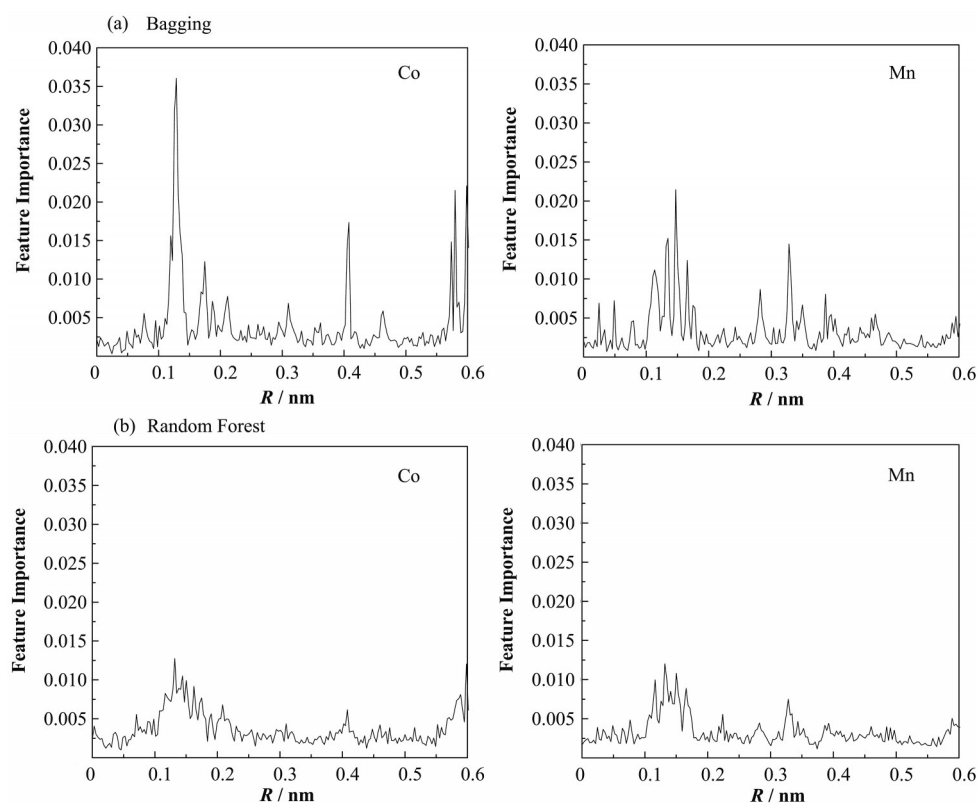


图8 特征重要度与 R 值的对应关系 (a) Co、Mn元素的袋装法基尼指数, (b) Co、Mn元素的随机森林基尼指数
Fig.8 Relationship between feature importance and R value (a) Gini index of bagging method for Co and Mn elements, (b) Gini index of random forest for Co and Mn elements

5 结语

本研究成功开发并验证了一种基于机器学习的方法,用于从EXAFS谱图中预测第四周期过渡金属元素的配位数。通过从Materials Project数据库中收集大规模EXAFS数据,并结合神经网络和决策树算法,我们的模型在配位数分类任务中取得了约70%的平均预测准确率。特别值得注意的是,神经网络模型在某些元素(如V)上表现尤为出色,准确率高达81.74%,展现了其处理复杂谱图数据的强大潜力。

研究结果表明,随机森林模型不仅在预测性能上与神经网络相当,还通过特征重要性分析揭示了EXAFS谱图的关键信息。具体来说,傅里叶变换后 R 空间中 $R < 0.3$ nm数据点的强度信息对配位数预测至关重要。这一发现与EXAFS理论中近程原子间相互作用主导谱图特征的观点一致,为模型的决策过程提供了理论支持和可解释性。

然而,研究也发现模型的预测准确率存在显著的元素依赖性。例如,Co元素的准确率较低(67.39%),这可能与数据库中配位数的分布多样性有关。此外,特征重要性分析显示,某些元素(如Co)的峰位置和强度参数的重要性低于预期,提示

可能存在未被充分利用的谱图特征。这一现象表明,当前的特征提取方法仍有改进空间。

未来研究方向应包括优化特征工程,探索更精细的谱图特征提取技术,以及尝试其他先进的机器学习算法,以提升模型在不同元素上的预测准确性和泛化能力。同时,结合物理模型与机器学习的混合方法也值得进一步研究,以增强EXAFS谱图解析的效率和可靠性。本研究的成果为材料局部结构的快速解析提供了新的工具和思路,具有重要的科学意义和应用前景。

作者贡献声明 曾海滔负责实验方案设计、数据收集与计算、稿件撰写;胡龙飞提出实验方案框架和参与稿件修改;姚涛负责整体实验方案指导和稿件修改。

参考文献

- 1 Terry J, Lau M L, Sun J T, *et al.* Analysis of extended X-ray absorption fine structure (EXAFS) data using artificial intelligence techniques[J]. Applied Surface Science, 2021, **547**: 149059. DOI: 10.1016/j.apsusc.2021.149059.
- 2 Tetef S, Govind N, Seidler G T. Unsupervised machine learning for unbiased chemical classification in X-ray

- absorption spectroscopy and X-ray emission spectroscopy [J]. *Physical Chemistry Chemical Physics*, 2021, **23**(41): 23586 – 23601.
- 3 Timoshenko J, Anspoks A, Cintins A, *et al.* Neural network approach for characterizing structural transformations by X-ray absorption fine structure spectroscopy[J]. *Physical Review Letters*, 2018, **120**(22): 225502. DOI: 10.1103/physrevlett.120.225502.
- 4 James G, Witten D, Hastie T, *et al.* *An introduction to statistical learning: with applications in Python*[M]. Cham: Springer International Publishing, 2023. DOI: 10.1007/978-3-031-38747-0.
- 5 Hoerl A E, Kennard R W. Ridge regression: biased estimation for nonorthogonal problems[J]. *Technometrics*, 2000, **42**(1): 80. DOI: 10.2307/1271436.
- 6 Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective[J]. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2011, **73**(3): 273 – 282. DOI: 10.1111/j.1467-9868.2011.00771.x.
- 7 Ouyang R H, Curtarolo S, Ahmetcik E, *et al.* SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates[J]. *Physical Review Materials*, 2018, **2**(8): 083802. DOI: 10.1103/physrevmaterials.2.083802.
- 8 Quinlan J R. Induction of decision trees[J]. *Machine Learning*, 1986, **1**(1): 81 – 106. DOI: 10.1007/BF00116251.
- 9 Opitz D, Maclin R. Popular ensemble methods: an empirical study[J]. *Journal of Artificial Intelligence Research*, 1999, **11**: 169 – 198. DOI: 10.1613/jair.614.
- 10 Breiman L. Bagging predictors[J]. *Machine Learning*, 1996, **24**(2): 123 – 140. DOI: 10.1007/BF00058655.
- 11 Ho T K. The random subspace method for constructing decision forests[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(8): 832 – 844. DOI: 10.1109/34.709601.
- 12 Freund Y, Schapire R E. *A decision-theoretic generalization of on-line learning and an application to boosting*[M]. *Computational Learning Theory*. Berlin, Heidelberg: Springer, 1995: 23 – 37. DOI: 10.1007/3-540-59119-2_166.
- 13 Torrisi S B, Carbone M R, Rohr B A, *et al.* Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships[J]. *NPJ Computational Materials*, 2020, **6**: 109. DOI: 10.1038/s41524-020-00376-6.
- 14 Sayers D E, Stern E A, Lytle F W. New technique for investigating noncrystalline structures: Fourier analysis of the extended X-ray: absorption fine structure[J]. *Physical Review Letters*, 1971, **27**(18): 1204 – 1207. DOI: 10.1103/physrevlett.27.1204.
- 15 Sun Z H, Liu Q H, Yao T, *et al.* X-ray absorption fine structure spectroscopy in nanomaterials[J]. *Science China Materials*, 2015, **58**(4): 313 – 341. DOI: 10.1007/s40843-015-0043-4.
- 16 杨旺, 林金如, 姚慧, 等. 针铁矿和赤铁矿吸附 As(V) 的 EXAFS 研究[J]. *核技术*, 2021, **44**(12): 120101. DOI: 10.11889/j.0253-3219.2021.hjs.44.120101.
YANG Wang, LIN Jinru, YAO Hui, *et al.* EXAFS study on arsenate(V) adsorption onto goethite and hematite[J]. *Nuclear Techniques*, 2021, **44**(12): 120101. DOI: 10.11889/j.0253-3219.2021.hjs.44.120101.
- 17 陈雨, 王建东, 浦世节, 等. 用于 X 射线吸收精细结构测量荧光气体探测器的设计与研制[J]. *核技术*, 2024, **47**(12): 120401. DOI: 10.11889/j.0253-3219.2024.hjs.47.120401.
CHEN Yu, WANG Jiandong, PU Shijie, *et al.* Design and development of fluorescence gas detector for X-ray absorption fine structure[J]. *Nuclear Techniques*, 2024, **47**(12): 120401. DOI: 10.11889/j.0253-3219.2024.hjs.47.120401.
- 18 Rehr J J, Kas J J, Vila F D, *et al.* Parameter-free calculations of X-ray spectra with FEFF9[J]. *Physical Chemistry Chemical Physics*, 2010, **12**(21): 5503 – 5513.
- 19 Joly Y. X-ray absorption near-edge structure calculations beyond the muffin-tin approximation[J]. *Physical Review B*, 2001, **63**(12): 125120. DOI: 10.1103/physrevb.63.125120.
- 20 Newville M. Larch: an analysis package for XAFS and related spectroscopies[J]. *Journal of Physics: Conference Series*, 2013, **430**: 012007. DOI: 10.1088/1742-6596/430/1/012007.
- 21 Pan H, Ganose A M, Horton M, *et al.* Benchmarking coordination number prediction algorithms on inorganic crystal structures[J]. *Inorganic Chemistry*, 2021, **60**(3): 1590 – 1603. DOI: 10.1021/acs.inorgchem.0c02996.
- 22 Brunner G O. A definition of coordination and its relevance in the structure types AlB₂ and NiAs[J]. *Acta Crystallographica Section A*, 1977, **33**(1): 226 – 227. DOI: 10.1107/s0567739477000461.
- 23 O'Keefe M, Brese N E. Atom sizes and bond lengths in molecules and crystals[J]. *Journal of the American*

- Chemical Society, 1991, **113**(9): 3226 – 3229. DOI: 10.1021/ja00009a002.
- 24 O'Keeffe M. A proposed rigorous definition of coordination number[J]. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 1979, **35**(5): 772 – 775. DOI: 10.1107/s0567739479001765.
 - 25 Voronoi G. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites[J]. Journal Für Die Reine und Angewandte Mathematik (Crelles Journal), 1908, **1908**(133): 97 – 102. DOI: 10.1515/crll.1908.133.97.
 - 26 Zimmermann N E R, Jain A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity [J]. RSC Advances, 2020, **10**(10): 6063 – 6081.
 - 27 Ong S P, Richards W D, Jain A, *et al.* Python materials genomics (pymatgen): a robust, open-source Python library for materials analysis[J]. Computational Materials Science, 2013, **68**: 314 – 319. DOI: 10.1016/j.commatsci.2012.10.028.
 - 28 Agarap A F M. Deep learning using rectified linear units (ReLU) [EB/OL]. [2018-05-22] <https://arxiv.org/abs/1803.08375>.
 - 29 Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California, USA, 2016: 785 – 794. DOI: 10.1145/2939672.2939785.
 - 30 Lee K L K, Gonzales C, Nassar M, *et al.* MatSciML: A broad, multi-task benchmark for solid-state materials modeling[EB/OL]. [2023-09-12] <https://arxiv.org/abs/2309.05934>.
 - 31 Chanussot L, Das A, Goyal S, *et al.* Open catalyst 2020 (OC20) dataset and community challenges[J]. ACS Catalysis, 2021, **11**(10): 6059 – 6072. DOI: 10.1021/acscatal.0c04525.
 - 32 Jain A, Ong S P, Hautier G, *et al.* Commentary: The materials project: a materials genome approach to accelerating materials innovation[J]. APL Materials, 2013, **1**: 011002. DOI: 10.1063/1.4812323.
 - 33 Kirklin S, Saal J E, Meredig B, *et al.* The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies[J]. NPJ Computational Materials, 2015, **1**: 15010. DOI: 10.1038/npjcompumats.2015.10.
 - 34 Ghigna P, Di Muri M, Spinolo G. Computer simulation approach to reliability and accuracy in EXAFS structural determinations[J]. Journal of Applied Crystallography, 2001, **34**(3): 325 – 329. DOI: 10.1107/s0021889801004745.
 - 35 Marcella N, Shimogawa R, Xiang Y C, *et al.* First shell EXAFS data analysis of nanocatalysts *via* neural networks [J]. Journal of Catalysis, 2025, **447**: 116145. DOI: 10.1016/j.jcat.2025.116145.
 - 36 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python[EB/OL]. [2018-07-15] <https://arxiv.org/abs/1201.0490>.