

基于机器学习的风化硅酸盐玻璃原成分预测及亚分类方法

王祉皓¹, 赵茗澈¹, 李智群¹, 郭 明¹, 肖琬玥¹, 刘志坚²

(1. 大连海事大学船舶电气工程学院, 辽宁 大连 116026; 2. 大连海事大学轮机工程学院, 辽宁 大连 116026)

摘 要: 玻璃在中国史料中早有记载, 但是由于长期存在名称和质地的混淆, 且近现代有关中国古代玻璃的研究起步较晚, 关于古代硅酸盐玻璃的风化和成分研究比较缺乏。以往研究古代玻璃器的著作, 多是从王朝更替的角度, 对文化交流、化学分析等方面研究玻璃的文化艺术形态及其自身的运行发展的规律, 较少有学者系统建立数学模型并使用智能算法定性定量开展风化硅酸盐玻璃原成分预测及亚分类方法研究。本工作以多组风化和未风化硅酸盐玻璃为研究对象, 提取其化学成分含量、纹饰和颜色等数据, 利用 Spearman 系数分析了纹饰、颜色和玻璃大类之间的相关性并研究影响表面风化的因素; 利用决策树进行大致分类, 然后用神经网络预测玻璃风化前主要化学成分的含量, 并总结硅酸盐玻璃的分类依据。此后通过 K-means 聚类建立分类模型: 确定最佳类别数, 进行亚类划分, 寻找铅钡玻璃和高钾玻璃的最优分类数量。研究结果表明, 只有玻璃类型对表面风化具有显著影响; 风化过程中参与度较高的化学成分为二氧化硅、氧化铝、氧化铅、氧化钡、氧化铅和五氧化二磷; 风化后, 铅钡玻璃二氧化硅含量明显下降, 氧化铅含量明显上升, 而高钾玻璃二氧化硅含量明显上升, 氧化钾氧化钙和氧化铝含量明显下降; 高钾玻璃分为 3 个亚类, 铅钡玻璃分为 4 个亚类。为后续利用机器学习研究古代硅酸盐玻璃的风化和成分提供了参考。

关键词: 硅酸盐玻璃; 决策树; 神经网络; K-means 聚类; Spearman 相关性

中图分类号: O756 文献标志码: A 文章编号: 0454-5648(2023)02-0416-11

网络出版时间: 2023-01-17



Prediction of Original Ingredients of Portland Glass and Research into Subclassification Methods Based on Machine Learning

WANG Zhihao¹, ZHAO Xingwei¹, LI Zhiqun¹, GUO Ming¹, XIAO Wanyue¹, LIU Zhijian²

(1. Marine Electrical Engineering College, Dalian Maritime University, Dalian 116026, Liaoning, China; 2. Marine Engineering College, Dalian Maritime University, Dalian 116026, Liaoning, China)

Abstract: Glass as a material has existed in China for a long time, but the related studies on ancient glass in China started relatively late due to the long-term confusion of name and texture, leading to a lack of research on the weathering and composition of ancient silicate glass. Some previous studies on ancient glass mainly discussed the artistic character and development laws of glass with respect to cultural exchange and chemical analysis from the perspective of dynastic succession. A few work established the related mathematical model and used the intelligent algorithm for qualitative quantification of weathering silicate glass original composition prediction and subclassification method. This paper was to use multiple groups of weathered and unweathered silicate glasses and collect/extract the data on their chemical composition content, ornamentation and color. The relations among the patterns, color, types of glass and surface weathering were analyzed by the Spearman coefficient. The decision tree for a rough classification and neural network to predict the main chemical composition of glass before its weathering was given, and the classification basis of silicate glass was summarized. Besides, the subcategorization at the optimal quantity of categories to conduct subclass classification was established, and a reasonable amount of barium glass and high potassium glass was selected. The results show that the type of glass has an influence on the surface weathering, and there are silicon dioxide, aluminum oxide, lead oxide, barium oxide and phosphorus pentoxide involved in the weathering process. Moreover, the amount of silicon dioxide decreases and lead oxide increases sharply in lead barium glass, while vice versa in high potassium glass after weathering.

Keywords: Portland glass; decision tree; neural network; K-means clustering; Spearman coefficient

收稿日期: 2022-11-15。 修订日期: 2022-12-13。

基金项目: 国家自然科学基金(51909019)。

第一作者: 王祉皓(2001—), 男, 本科生。

通信作者: 刘志坚(1986—), 男, 硕士, 副教授。

Received date: 2022-11-15. Revised date: 2022-12-13.

First author: WANG Zhihao (2001—), male, Undergraduate.

E-mail: wangzhihao@dlnu.edu.cn

Correspondent author: LIU Zhijian (1986—), male, Master, Associate Professor.

E-mail: liuzhijian@dlnu.edu.cn

玻璃,几千年来从稀有走向普通,再到现如今在各行各业中发挥重要的作用。这其中,又以硅酸盐玻璃数量众多,应用广泛。本工作希望通过研究古代硅酸盐玻璃的风化与亚分类问题,为当代硅酸盐玻璃预防风化、应用和数字化研究做一些先行性研究。

首先是风化,古代玻璃极易受埋藏环境的影响而风化。周良知^[1]应用近代研究方法,研究影响硅酸盐玻璃风化的主要因素。王承遇等^[2]对风化作用的机理所进行的深入研究,对于成分预测工作有一定的借鉴意义。其次是分类。虽有赵凤燕等^[3]通过 pXRF 对玻璃器的化学成分进行无损分析来进行分类,但是现有的化学研究手段还不能准确按照玻璃的成分进行较为精准合理的分类。因此,考虑引入机器学习来通过使用分类预测模型和智能算法来将实际的化学问题抽象为数理模型来解决,以免去分析时实体样本和高精度分析仪器的束缚。虽然智能算法近年来广泛应用于材料领域的科学研究,如冯百龄^[4]研究的中国出土古代玻璃珠数据库建设与应用,张丽艳等^[5]使用 7 种玻璃成分性质模拟方法总结了各模拟方法的主要理论依据、模拟过程及应用现状和李建刚^[6]基于深度学习做的玻璃缺陷检测,但是国内外在使用机器学习研究古代硅酸盐玻璃的风化与亚分类问题方面尚且空白。

机器学习可以辅助科学家从不同尺度、维度深入认识材料的机理特征,理解问题的本质^[7]。本研究过程中主要使用的算法是神经网络、决策树和

K-means 聚类。其中神经网络主要用来解决风化前后化学成分含量变化问题,决策树主要用来解决玻璃的大致分类问题^[8]; K-means 聚类作为一种基于划分的聚类算法^[9],是从无标记数据集中获取信息和知识的重要手段。由于其具有适用范围广泛和算法伸缩性强的优点, K-means 聚类算法被广泛的应用到聚类分析领域,因为其良好的分类划分效果,可以满足大多数分类要求^[10]。

1 机器学习

1.1 机器学习概述

机器学习,涉及概率论、统计学、逼近论、凸分析、算法复杂度等多门学科,是专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能的一门多领域交叉学科^[11]。目前,机器学习在基于知识的系统中、自然语言理解、非单调推理、机器视觉、模式识别等许多领域都得到了广泛应用,计算化学也是飞速发展^[12]。模型目前已被广泛应用于建立机器学习模型的输入量有多种表述方法,如变量、特征、描述符等,不同领域的表述方法也不尽相同。

根据学习方式的不同,机器学习可分为有监督学习^[13]、无监督学习^[14]、自监督学习^[15]、半监督学习、主动学习和强化学习等。各类机器学习方法近年来广泛应用于材料研究领域,具体应用可简要概括如图 1 所示。

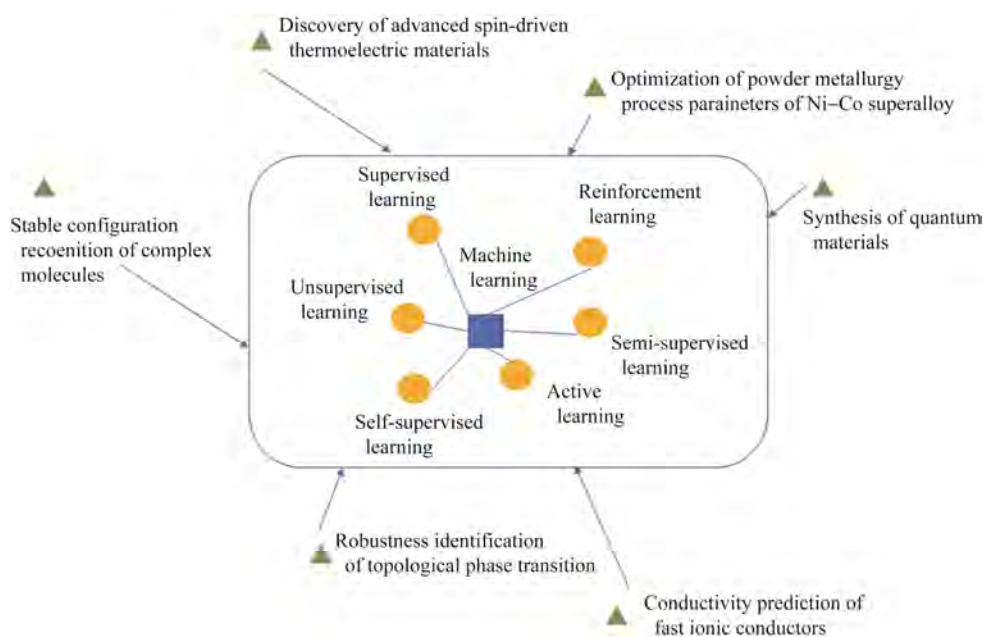


图 1 各类机器学习方法在材料领域科学研究应用

Fig. 1 Application of various machine learning methods in scientific research in the field of materials

上图各算法中, 决策树和神经网络算法属于有监督学习, *K*-means 聚类算法属于无监督学习。本研究将利用这 2 种算法建立模型。

1.2 决策树样本集

决策树算法具有算法原理层面的随机性和集成学习的优势, 在较高维度的问题特征处理方面具有良好的稳定性与敏感性。因此本工作将基于决策树算法构建粗分类模型。构成该部分模型的关键是样本集的建立以及抽样, 形成相应的子决策树, 最终的输出结果由各个子决策树输出结果用相对多数投票法决定^[10]。

子决策树的数量 m 对于整个算法的计算效率有着极大的影响^[16], 随着子决策树的数量增大, 分类精度会更好, m 值的过大会导致算法的运行效率下降, 但是如果 m 值过小, 则分类精度将可能下降。

1.3 决策树分类的实现

根据 GiNi 系数对决策树的数据进行处理, 形成训练集。然后根据训练集中的元素建立集合表示为 $\{c_1, c_2, \dots, c_N\}$, 为方便讨论, 这里将 h_i 在样本 x 的预测输出表示为一个 N 维向量 $(h_i^1(x), h_i^2(x), \dots, h_i^N(x))^T$, 其中 $h_i(x)$ 表示 h_i 在类别 c_j 上的输出, 建立投票表达式如下:

$$H(x) = \begin{cases} c_j, & \sum_{i=1}^T h_i^j(x) > 0.5 \sum_{k=1}^N \sum_{i=1}^T h_i^k(x) \\ \text{Reject, 其他} \end{cases} \quad (1)$$

其中: $H(x)$ 为分类的结果输出, 即若某个标记得票数过半数, 则应判断为该类别, 否则拒绝判断。

1.4 神经网络预测的实现

神经网络算法具有自学习功能和高速寻找优化解的能力。因此将基于神经网络算法构建风化预测模型。在样本集抽样出合理的数据集之后, 应将多组数据集进行随机对应, 形成训练集, 作为神经网络的输入数据集的特征变量, 同时将特征变量作为输入信号代入式(2), 进行加权平均计算:

$$y' = W_1 X_1 + W_2 X_2 + \dots + W_n X_n \quad (2)$$

其中: X_i 为第 i 个变量的输入信号; W_i 为输入信号 X_i 权值; y' 为加权后的决策变量。

为方便进行后续数据处理, 引入激活函数 ψ 来归一化处理数据,

$$\psi(y') = \frac{1}{1 + e^{-y'}} \quad (3)$$

此式引入激活函数做归一化处理, 对自变量进行修改, 因此该式中各变量仅有数值含义。对加权平均结果进行非线性变换, 选用如式(2)所示的 Sigmoid 函数做激活函数, 将原取值范围 $(-\infty, +\infty)$

转换到 $(0, 1)$ 。

上述为单层神经网络, 适用于数据量较小的样本集, 若样本集数据量较大, 可采用多层神经网络。

1.5 肘部原则

K-means 在进行分类计算时, 具有极佳的计算简易性, 并且降低了计算的复杂度, 但是在进行 *K*-means 聚类计算之前, 最为重要的便是分类数 K 值的选取^[17]。

肘部法则对于 *K*-means 算法的 K 值确定起到指导作用^[18], 该法则的计算原理是成本函数, 成本函数是类别畸变程度之和, 每个类的畸变程度等于每个变量点到其类别中心的位置距离平方和, 若内部的成员彼此间越紧凑, 则类的畸变程度越小; 反之, 若该类内部的成员彼此间越分散, 则类的畸变程度越大。随着值的增大, 平均畸变程度会减小; 每个类包含的样本数会减少, 于是样本会更靠近其中心。但是, 随着值继续增大, 平均畸变程度的改善效果会不断减低。值增大过程中, 畸变程度的改善效果下降幅度最大的位置对应的值就是肘部, 即下降程度突然变缓时即认为最佳的 K 值^[18]。

1.6 *K* 聚类算法概述

K-means 聚类算法属于一种无监督的机器学习算法, 是将归类相似的对象归到同一个簇中, 簇内对象越相似, 聚类的效果越好。首先, 确定 K 个初始点作为质心。在此基础上, 按照式(4)为每个点找到距其最近的质心, 并将其分配给质心所对应的簇。

$$D_i = |S_i - S_0| \quad (4)$$

其中: S_i 代表第 i 个随机点; S_0 为质心点; D_i 为第 i 个随机点与质心的距离。

在完成上述步骤后, 将每个簇的质心更新为该簇内所有点的平均值。按照上述方式反复进行迭代, 直到所有的数据点的簇分配结果不再改变为止。

K-means 中输入的点一般为多个特征变量组成的多维点坐标, 输出分类结果。

1.7 轮廓系数

轮廓系数是聚类效果好坏的一种评价方式^[17], 它结合内聚度和分离度 2 种因素, 可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。轮廓系数 $S(i)$ 公式如下:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (5)$$

其中: $a(i)$ 为样本点的内聚度, 也即类内距离; $b(i)$ 为类间距离。

$a(i)$ 计算公式如下:

$$a(i) = \frac{1}{n-1} \sum_{j \neq i}^n dis(i, j) \quad (6)$$

其中: j 为样本 i 在同一个类内的其他样本点; d 为 i 与 j 的距离。所以 $a(i)$ 越小说明该聚类越紧密。 $b(i)$ 的计算方式与 $a(i)$ 类似。需要注意的是, 遍历其他类得到多个值 $\{b1(i), b2(i), b3(i), \dots, bm(i)\}$ 从中选择最小的值作为最终的结果。所以原轮廓系数 $S(i)$ 可化简为式(7):

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & a(i) < b(i) \\ 0 & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & a(i) > b(i) \end{cases} \quad (7)$$

根据式(7)可发现: 当 $a(i) < b(i)$ 时, 即类内距离小于类间距离, 则聚类结果更紧凑; S 的值会趋近 1, 越趋近于 1 代表轮廓越明显, 分类效果越好; 相反, 当 $a(i) > b(i)$ 时, 类内距离大于类间距离, 说明聚类结果很松散, S 的值会趋近 -1, 越趋近于 -1 则聚类的效果越差, 需要考虑重新进行分类。

2 模型建立

2.1 数据获取与预处理

通过查阅资料^[19]整理出古代风化前后硅酸盐玻璃的相关数据。由于存在数据缺失, 即在该玻璃样品中该化学成分未检测到, 考虑使用 0 值进行补全。此外, 由于检测手段等原因可能导致其成分比例的累加和非 100% 的情况^[20], 本研究将成分比例累加和介于 85%~105% 的数据视为有效数据, 其余的视为无效数据, 予以剔除。另外需要注意的是, 该数据集检测到二氧化硫的采样点过少, 因此在后续分析中对该化学成分不予考虑。

进行数据基础预处理后, 运用方差滤波进行降维, 认为方差小于预设值的变量对分类及成分含量结果关联度较小, 因此予以剔除, 从而实现特征变量的降维。

2.2 风化预测模型

首先研究硅酸盐玻璃表面风化与其玻璃类型、纹饰和颜色的关系。对于样本容量为 n 的样本, n 个原始数据被转换成等级数据, Spearman 相关系数 ρ 的计算式^[21]为:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (8)$$

其中: x_i 与 y_i 分别为硅酸盐玻璃表面风化、玻璃类型、纹饰和颜色 4 个变量的任意 2 个变量的数据集, $\bar{x}$, $\bar{y}$ 为两者数据集的平均值。

由于 Spearman 相关系数表明独立变量和依赖变量的相关方向, 所以 ρ 为负意味着当独立变量增加时, 依赖变量趋向于减少; ρ 为正则相反; ρ 在绝对值上增加则意味着独立变量和依赖变量越来越接近完全的单调相关^[21-22]。

根据前述分析, 只对参与风化作用多的化学物质进行预测, 分别对高钾玻璃和铅钡玻璃进行预测。以铅钡玻璃为例, 将其分为已风化和未风化 2 个数据集作为输入信号(在数据量偏少的情况下可以将 2 个数据集按照化学含量一致原则进行随机对应, 形成训练集)。对输入的信号进行加权平均, 对激活函数 Ψ 的结果进行加权平均结果进行非线性变换。最终便可以得到硅酸盐玻璃未风化时的化学物质含量。

2.3 其他类别的机器学习模型

在机器学习中, 依照学习类型进行分类, 可以分为有监督学习和无监督学习。有监督学习可以从训练数据中建立学习模型, 然后根据模型推断出新的样本, 目的是通过机器学习建立一个从输入数据到目标数据的映射。比起有监督学习, 无监督学习能够自主解释数据并有可能发现一些新的模式。

依照模型类别进行分类, 可以分为线性模型和非线性模型。其中线性模型是机器学习中最简单的模型之一^[23]。其目的是建立输入数据到输出数据之间的线性关系, 常常无法解决复杂性过高的问题, 但是由于可解释性强, 使得其在机制解释, 性能影响因素的研究中表现极其突出。比起线性模型, 非线性模型则可以找出二次以上未知量之间的关系, 但是运算相对来说更加复杂。

2.4 玻璃分类模型

为方便后续处理, 将硅酸盐玻璃中的各化学成分编号, 其对应关系如表 1 所示:

表 1 化学成分与编号对照表

Table 1 Comparison table of chemical composition and serial number

Sample No.	Component	Sample No.	Component
0	SiO ₂	7	CuO
1	Na ₂ O	8	PbO
2	K ₂ O	9	BaO
3	CaO	10	P ₂ O ₅
4	MgO	11	SrO
5	Al ₂ O ₃	12	SnO ₂
6	Fe ₂ O ₃	13	SO ₂

利用神经网络算法, 代入 Python 进行计算得到其分类规律, 找出风化的主要影响因素。在该影响因素的基础上对硅酸盐玻璃进行亚分类, 利用 K-means 聚类算法进行分类。

根据肘部原则, 通过计算最小化点到聚类中心的距离 $D_{\min}$, 来确定的类别数。然后将最小化点与均值的距离作为因变量, 类别数作为自变量, 在绘制图像之后, 进行分析图像, 便可以得到最佳的分类数。在得到最佳的分类数之后, 代入 K-means 聚类算法当中, 便可对硅酸盐玻璃进行亚分类, 在得到亚分类之后进一步分析, 便可得到分类依据。

3 结果及分析

3.1 风化预测模型结论

首先将已有数据代入式(8)进行计算, 得到图 2。图 2 中蓝色代表正相关, 浅蓝色代表负相关, 颜色越深代表相关性越强, 颜色越浅, 代表数值越小。由图 2 可得出是否风化只与玻璃种类显著性相关, 与纹饰和颜色无关。这一发现证明风化现象发生应该主要进行化学分化, 而对于硅酸盐玻璃而言物理风化等现象相对较少。

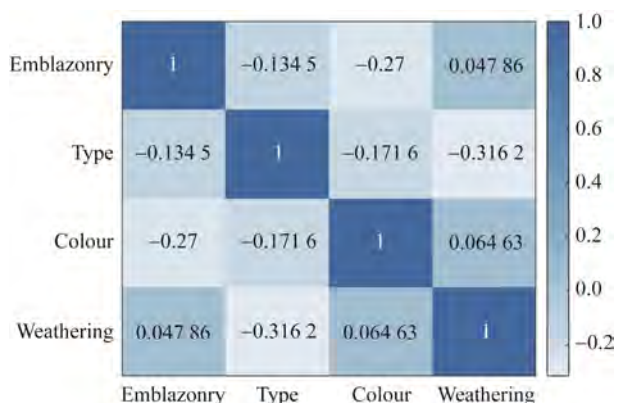


图 2 纹饰、颜色、类型和表面风化的热值图

Fig. 2 Calorific value map of texture, color, type and surface weathering

纹饰和颜色对是否风化没有显著性关系, 因此分析时可不予考虑。根据方差滤波分析可知: 对于铅钡玻璃而言, 二氧化硅、氧化钡、氧化铅、五氧化二磷、氧化铜和氧化铝为风化主要影响成分; 对于高钾玻璃而言, 二氧化硅、氧化钾、氧化钙和氧化铝为风化主要影响成分。即这几部分物质含量的变化对于研究者确定硅酸盐玻璃是否风化起了决定性作用。

分别做出铅钡玻璃和高钾玻璃的风化与未风化的化学成分柱状图, 如图 3、图 4 所示。

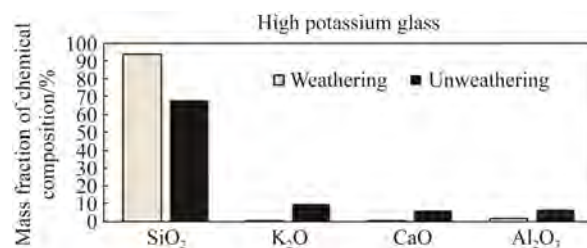


图 3 铅钡玻璃的风化主要影响因素柱状图

Fig. 3 Histogram of main factors affecting the weathering of lead barium glass

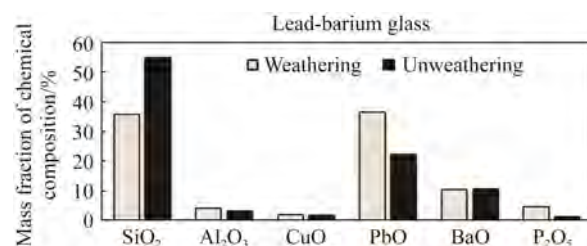


图 4 高钾玻璃的风化主要影响因素柱状图

Fig. 4 Histogram of main factors affecting weathering of high potassium glass

由图 3、图 4 可以得到统计规律: 铅钡玻璃风化后, 二氧化硅的质量分数(下同)明显下降, 氧化铁含量下降不明显, 氧化钡基本保持不变, 氧化铝含量略有升高, 氧化铅和五氧化二磷含量显著提升; 高钾玻璃风化后, 二氧化硅的含量明显上升, 氧化钾、氧化钙和氧化铝的含量显著下降。对于神经网络算法来说, 训练集的选取十分重要, 本工作在选取训练集前对数据集^[12]进行详细分析发现: 1) 严重风化点处的风化特征最突出, 因此认为选取训练集时可以将严重风化点作为固定选取对象; 2) 部分硅酸盐玻璃大多数表面发生了风化, 但是采样时恰好采集到了未风化点处的化学成分含量, 则其本身具有强对比度, 在预测时具有较高的参考价值。

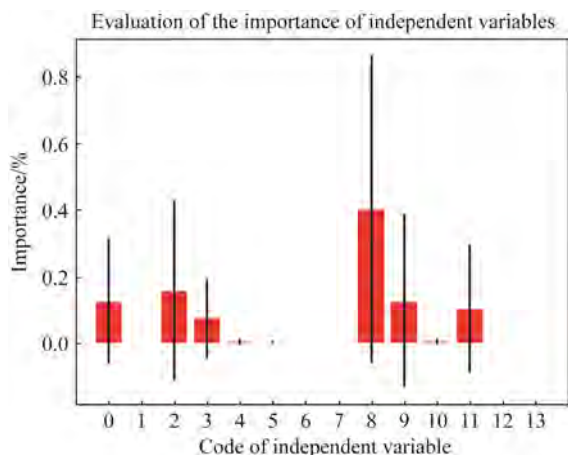
为提高预测精度, 选取训练集时将上述 2 种采样点作为固定选取对象, 与其他已风化采样点共同组成训练集, 然后以该结果预测已风化硅酸盐玻璃风化前的各化学成分含量, 按照此模型便可以得到具体预测结果。类似地, 对高钾玻璃进行预测, 也可以得到具体预测结果。

在进行预测结论的数据分析时, 观测模型的预测数据, 发现跟实际成分差异较小, 在利用随机数进行调整输入数据的数值大小之后, 可以发现在误差允许范围内, 模型具有良好的鲁棒性^[24–25]。

3.2 玻璃分类模型结论

首先利用神经网络算法, 计算得出各化学成分

的影响程度, 并且将该规律绘制图 5。由图 5 可知: 均值越大, 相对标准差越大, 意味着该化学成分对分类依据的影响程度越大。结合求解结果, 可得到分类规律为: 分类时各化学成分优先级从高到低依次为氧化铅, 氧化钾, 氧化钡, 二氧化硅, 五氧化二磷, 氧化铝。



Note: Top four serial numbers are 8-lead oxide, 2-potassium oxide, 9-barium oxide, 0-silicon dioxide.

图 5 高钾玻璃与铅钡玻璃影响程度汇总图

Fig. 5 Histogram of main factors affecting weathering of high potassium glass

根据前述模型代入铅钡玻璃的数据, 利用肘部原则得到图 6 和图 7。如图 7 所示, 建立二维坐标, 利用其横纵坐标准确描述不同特征点的空间位置。由图 6 可知, 同时参照表 2 以及古代对硅酸盐玻璃种类的划分依据^[26], 进行亚类命名: 1 类的硅含量较高, 命名为高硅-铅钡玻璃; 2 类的铅含量较高, 命名为高铅-铅钡玻璃; 3 类的铝含量较低, 命名为

低铝-铅钡玻璃; 4 类没有明显特征, 命名为标准-铅钡玻璃。具体划分方法及划分结果见图 7。

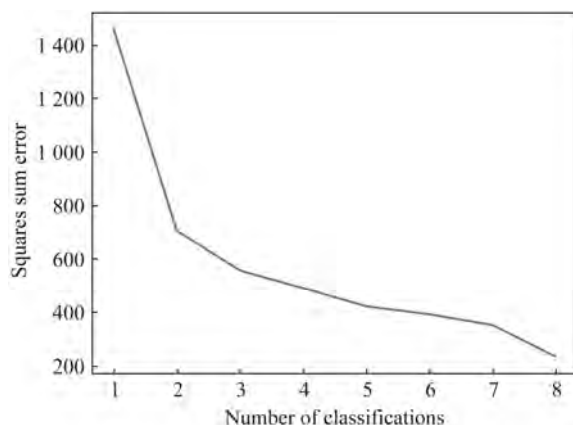
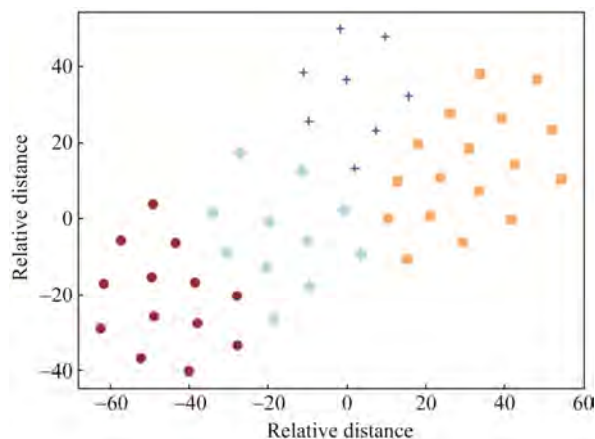


图 6 铅钡玻璃肘部原则折线图

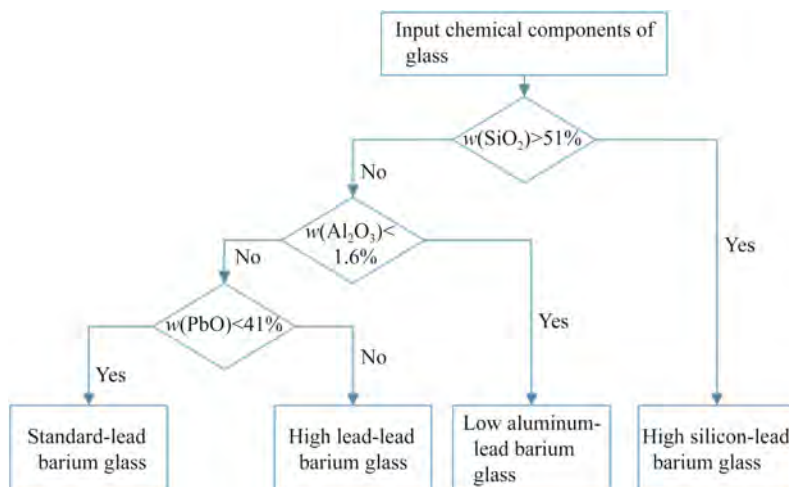
Fig. 6 Line chart of lead-barium glass elbow principle



Note: + is High lead-lead barium glass, ● is High silicon-lead barium glass, ■ is Low aluminum-lead barium glass, ◆ is Standard-lead barium glass.

图 7 铅钡玻璃亚分类示意图

Fig. 7 Schematic diagram of subclassification of lead-barium glass



w is mass fraction.

图 8 铅钡玻璃亚分类标准流程图

Fig. 8 Flow chart of sub classification standard for lead barium glass

类似地, 对高钾玻璃进行亚分类, 其肘部法则显示如图 9 所示。由图 9 可知: 2 类和 3 类处出现尖点, 即此后直线斜率明显减小, 但是 1 类 2 类均位于陡降区, 3 类 4 类位于缓降区。考虑到数据总量较少, 决定分为 3 个亚类。此后代入分类模型, 经过距离计算后分类, 大致分类示意图如图 10 所示: 建立二维坐标, 利用其横纵坐标准确描述不同特征点的空间位置。

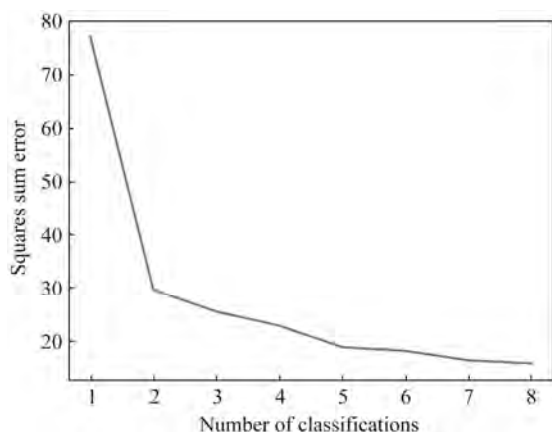
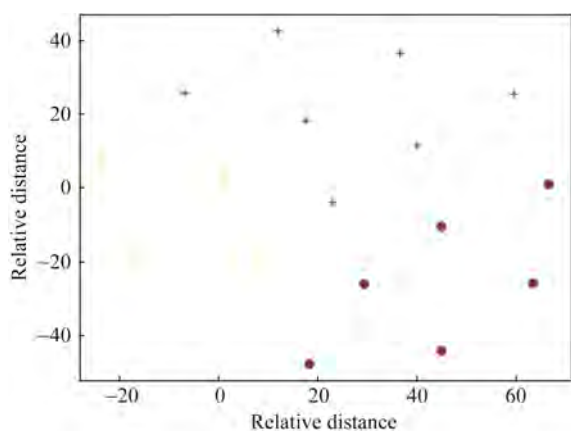


图 9 高钾玻璃肘部原则折线图

Fig. 9 Line diagram of high potassium glass elbow principles



Note: + is K-high potassium glass, ● is Si-high potassium glass, ■ is Al-high potassium glass.

图 10 高钾玻璃分类示意图

Fig. 10 Schematic diagram of subclassification of high potassium glass

按照相同的方式对高钾玻璃进行亚类命名: 分类一的硅含量较高, 将其命名为 Si-高钾玻璃; 分类二的铝含量较高, 将其命名为 Al-高钾玻璃; 分类三的钾含量较高, 将其命名为标准-高钾玻璃。具体分类标准见图 11。

根据亚分类的元素分类种类不同, 分析发现不管是高钾玻璃还是铅钡玻璃, 亚分类的分类标准均

与预测模型的主要影响化学物质的呈现对应关系, 同时赵凤燕等在实地采样研究中也应用了类似的分类思路^[3], 进一步证明二者结论的可靠性。

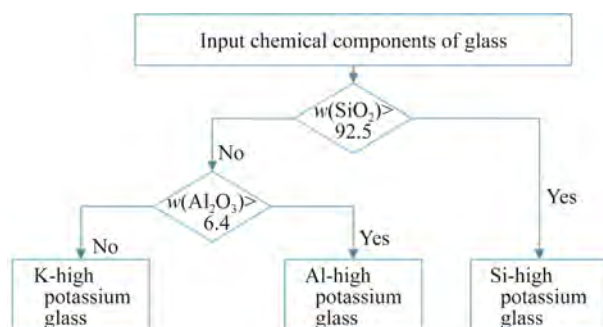


图 11 高钾玻璃亚分类标准流程图

Fig. 11 Flow chart of subclassification and classification standards for high potassium glass

3.3 模型合理性

使用了 *K-means* 聚类, 为分析模型的合理性, 利用轮廓系数作为聚类性能的评估指标。轮廓系数取值范围为 $[-1, 1]$, 取值越接近 1 则说明聚类性能越好, 相反, 取值越接近-1 则说明聚类性能越差。

以铅钡玻璃为例, 根据数据总量, 将分类条件划分成 2~40 类, 得到图 12。由图 12 可知, 虽然宏观来看分成 2 类时出现最大值, 但是分成 2 类不符合分类标准, 综合考虑之下, 选择出现局部最大值的地方, 即分成 4 类时, 该图像直观的说明了分类结果具有合理性。

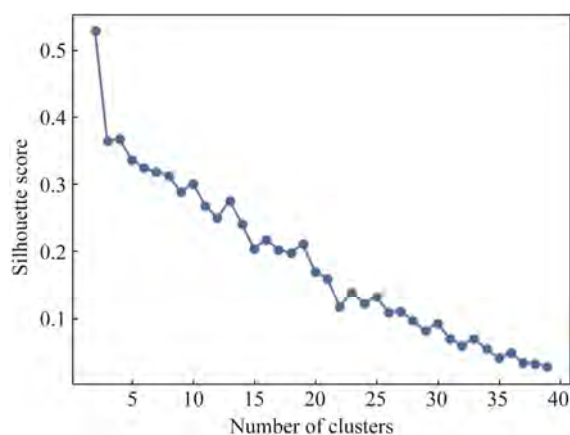


图 12 铅钡玻璃轮廓系数折线图

Fig. 12 Line chart of lead barium glass profile coefficient

类似地, 可以建立高钾玻璃的轮廓系数图(图 13), 全局最大值为分类点取在 3 时, 完全符合 *K-means* 聚类的分类情况, 基于上述对轮廓系数的分析, 该模型具有很好的合理性。

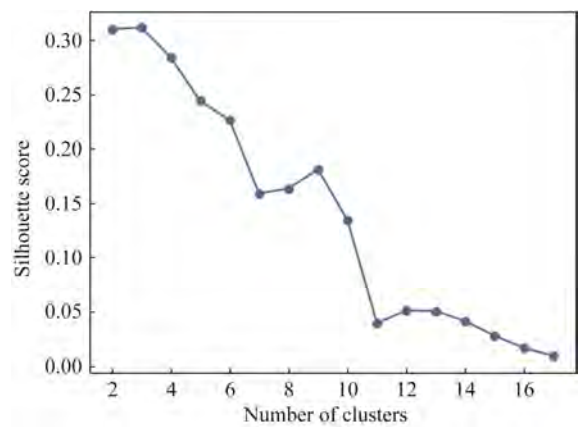


图 13 高钾玻璃轮廓系数折线图

Fig. 13 Line chart of high potassium glass profile coefficient

3.4 模型准确性

对于风化预测模型而言，本工作采用的单层神

经网络适用于数据量较少的样本集。风化预测模型的神经网络预测结果如表 2 和表 3 所示。若样本集数据量较多，则需选用多层神经网络算法。需注意算法与样本规模的匹配度：若算法与样本规模不匹配，则极有可能出现拟合不当。此外，使用线性模型进行预测，分别用 R^2 检验 2 种机器学习方法的结果准确性，如表 4 和表 5 所示。

对比表 4 和表 5 中的数据可知，与线性模型相比，神经网络算法的结果精确度更高，更适用于本次研究的数据集。神经网络算法具有以下优点：计算简单，易于理解，可解释性强；比较适合处理有缺失属性的样本；能够处理不相关的特征；在相对短的时间内能够对大型数据源做出可行且效果良好的结果。

表 2 铅钡玻璃风化前质量分数化学成分预测结果

Table 2 Prediction results of mass fraction of chemical composition of lead barium glass before weathering															w/%
Sample No.	Surface weathering	SiO ₂	Na ₂ O	K ₂ O	CaO	MgO	Al ₂ O ₃	Fe ₂ O ₃	CuO	PbO	BaO	P ₂ O ₅	SrO	SnO ₂	SO ₂
A1	Unweathering	78.45	0.00	0.00	6.08	1.86	7.23	2.15	2.11	0.00	0.00	1.06	0.03	0.00	0.51
A2	Weathering	37.75	0.00	0.00	7.63	0.00	2.33	0.00	0.00	34.30	0.00	14.27	0.00	0.00	0.00
A3	Unweathering	31.95	0.00	1.36	7.19	0.81	2.93	7.06	0.21	39.58	4.69	2.68	0.52	0.00	0.00
A4	Unweathering	35.47	0.00	0.79	2.89	1.05	7.07	6.45	0.96	24.28	8.31	8.45	0.28	0.00	0.00
A5	Weathering	64.29	1.20	0.37	1.64	2.34	12.75	0.81	0.94	12.23	2.16	0.19	0.21	0.49	0.00
A6	Weathering	93.17	0.00	1.35	0.64	0.21	1.52	0.27	1.73	0.00	0.00	0.21	0.00	0.00	0.00
A7	Weathering	90.83	0.00	0.98	1.12	0.00	5.06	0.24	1.17	0.00	0.00	0.13	0.00	0.00	0.11
A8	Unweathering	51.12	0.00	0.23	0.89	0.00	2.12	0.00	9.01	21.24	11.34	1.46	0.31	0.00	2.26

表 3 高钾玻璃风化前质量分数化学成分预测结果

Table 3 Prediction results of mass fraction of chemical composition of high potassium glass before weathering															w/%
Sample No.	Mass fraction chemical composition														
	SiO ₂	K ₂ O	CaO	Al ₂ O ₃	Na ₂ O	MgO	Fe ₂ O ₃	CuO	PbO	BaO	P ₂ O ₅	SrO	SnO ₂	SO ₂	
7	76.26	11.73	4.58	7.08	0	0.54	0.2	1.54	0	0	0.36	0	0	0	
9	77.91	11.89	4.04	7.14	0	0	0.17	3.24	0	0	0.61	0	0	0	
10	78.96	11.99	3.64	7.17	0	0	0.32	1.55	0	0	0.35	0	0	0	
12	77.79	11.88	4.01	7.13	0	0	0.26	0.84	0	0	0	0	0	0	
22	77.85	12.16	5.33	7.42	0	0	0.29	1.65	0	0	0.15	0	0	0	
27	76.56	11.88	4.84	7.21	0	0.64	0.35	0.55	0	0	0.21	0	0	0	

表 4 铅钡玻璃 2 种机器学习方法 R^2 的比较表

Table 4 Comparison table of two machine learning methods R^2 for lead barium glass

Component	Algorithm	
	Linear	Neural network
SiO ₂	-7.221 5	0.965 5
BaO	0.808 1	0.998 0
Al ₂ O ₃	-0.061 0	0.319 5
CuO	0.688 4	0.921 2
P ₂ O ₅	-0.084 1	0.679 6
PbO	-0.157 8	0.997 4

在前序工作中已经得到了高钾玻璃和铅钡玻璃的分类标准，可以此为依据对未分类硅酸盐玻璃进

行种类预测，寻找到一批新的数据如表 6^[19]，带入预测模型，得到表 7 数据。

表 5 高钾玻璃的 2 种机器学习方法 R^2 的比较表

Table 5 Comparison table of two machine learning methods R^2 for high potassium glass

Component	Algorithm	
	Linear	Neural network
SiO ₂	-22.182 4	0.947 2
K ₂ O	0.981 4	0.999 2
Al ₂ O ₃	0.033 2	0.003 3
CaO	0.764 8	0.997 8

表 6 硅酸盐玻璃的化学成分含量预测
Table 6 Chemical composition content of silicate glass to be predicted

w/%

Sample No.	Mass fraction of chemical composition													
	SiO ₂	Al ₂ O ₃	CuO	PbO	BaO	P ₂ O ₅	Na ₂ O	K ₂ O	CaO	MgO	Fe ₂ O ₃	SrO	SnO ₂	SO ₂
2	50.26	2.80	0.19	24.60	5.49	0.24	0	1.05	2.34	1.18	1.86	0.19	0	0
8	64.30	2.05	1.86	21.38	10.64	0	0	0	1.48	0	0	0.37	0	2.58
11	53.23	4.08	2.30	21.10	9.35	1.47	0	0.21	3.51	0.71	0	0.37	0	0
19	51.30	4.38	0.70	24.52	9.09	0	0	0	2.93	0.59	1.33	0.19	0	0
26	65.63	2.51	1.76	21.57	10.66	0.11	0	0	1.44	0	0	0.45	0	1.96
34	57.86	4.21	0.65	22.72	6.81	0.53	0	0.25	0.78	0	0.47	0.22	0	0
36	57.07	4.34	0.41	21.79	5.21	1.04	2.22	0.14	0.37	0	0.32	0.22	0	0
38	57.55	3.28	0.71	23.35	7.20	0.32	1.38	0	0.68	0	0.29	0.41	0	0
39	59.87	3.33	2.70	27.25	10.24	1.30	0	0	1.11	0	0	0.61	0	0
40	58.36	3.95	2.63	30.54	10.96	0.87	0	0	1.87	0	0.19	0.68	0	0
41	48.38	1.88	0.65	22.99	8.07	0	0	0.44	4.96	2.73	1.79	0.47	0	0
48	48.76	5.30	5.79	22.69	8.16	5.12	0.80	0.32	2.82	1.54	1.03	0.25	1.31	0
49	47.75	2.58	0	22.00	7.13	0.11	0	0	4.58	1.47	2.74	0.46	0	0
50	52.01	2.65	0.88	23.17	8.53	0	0	0	3.19	0.47	0.33	0.66	0	0
52	53.63	2.79	1.03	23.53	8.32	0.65	1.22	0	2.27	0.55	0.23	0.44	0	0
54	54.23	2.43	1.27	26.41	9.23	0.18	0	0.32	3.19	1.28	0	0.88	0	0
56	56.11	2.97	0.20	21.83	7.09	0.45	0	0	1.21	0	0	0	0	0
57	57.10	3.79	0.51	22.37	8.58	0	0	0	1.31	0	0	0	0	0
58	51.86	3.99	0.54	23.41	8.46	0.24	0	0.34	3.49	0.79	0.86	0.24	0	0
08 (Serious)	56.80	2.36	0.62	22.40	3.81	0.38	0	0	3.19	0	0	0.53	0	15.03
26 (Serious)	58.88	3.09	0.28	21.81	3.11	1.42	0	0.40	3.01	0	0	0.62	0	15.95
43 (1)	51.58	5.67	2.79	27.05	11.18	0.06	0	0	5.24	0.89	0.76	0.64	0	0
43 (2)	47.82	2.95	0.38	25.20	8.75	0.05	0	0	6.40	0.95	1.39	0.47	0	0
51 (1)	49.40	2.49	0.18	22.84	7.82	0	0	0	3.58	1.19	1.19	0.39	0.47	0
51 (2)	46.76	2.74	0.36	25.61	8.64	0.23	0	0	5.13	1.45	0.42	0	0	0
54 (Serious)	49.43	2.27	0.72	30.31	9.07	0.22	0	0	0	1.11	0	1.12	0	0

表 7 预测分类结果汇总表
Table 7 Summary of forecast classification results

Code of artifact	A1	A2	A3	A4	A5	A6	A7	A8
Result	K-high potassium glass	Lead-barium glass	Lead-barium glass	Lead-barium glass	Lead-barium glass	K-high potassium glass	K-high potassium glass	Lead-barium glass

观察数据, 并和实际情况进行对照, 发现并没有出现差错, 故该模型可以进行准确而快速的预测玻璃的种类。

4 结论

机器学习本质是一个抓取变量特性的过程, 化学, 尤其材料化学, 则是一个需要分析变量的学科。未来的趋势是工业互联网和数字孪生^[27]。机器学习提供了从因直接到果的捷径, 因此有些研究可以简化中间步骤, 从而加快研究速度。虽然机器学习在实际应用中仍然面临不少的挑战^[28], 但是在过去的 10 年中, 计算化学领域一直在增长, 近年来它已经成为优化和预测化学研究的非常强大的工具^[29]。

本工作利用机器学习中的决策树算法、神经网络算法和 *K-means* 聚类算法分别进行模型建立, 其中粗分类模型能够迅速准确地对未知种类玻璃进行粗分类, 风化预测模型能够较为准确的预测风化作用影响下的原本化学成分含量, 具有较为普遍的实用性。而亚分类模型能够准确的根据玻璃的各化学成分含量进行分类, 得到了较为理想的细致分类结果, 对于硅酸盐玻璃的分类具有较好的参考价值。通过引入实际玻璃的成分, 并且对比分析 *R*² 等衡量参数, 发现通过模型预测的结果, 与实际差异不大。因此可以认为模型的准确性可以达到实际分析水平, 具有一定的普适性, 对于机器学习在相关领域的引进中起到了一定的推进作用。

在硅酸盐玻璃的有关研究当中引入机器学习算法,得到了引人瞩目的进展,在建立预测模型当中,通过处理有效的实验数据,可以得到数据之间的关系,进一步分析内在关联性,本文便利用机器学习找到了风化与未风化关系的核心影响因素。以此类推,如果数据更加丰富,可以利用机器学习算法得到更多的关联性,届时与实验技术融合,进一步进行实验开发,可以更加有效地提高材料的可利用性、材料实验的效率和智能化。

在亚分类模型建立当中,对数据处理时预先进行了颜色可视化处理,凸显效果明显,提高了数据分析的效率,并且大量使用大数据分析,最大化地开发了数据的功能,发挥了数据的作用,使结论更加精准。但与此同时由于采集数据较少在数据量较少的组别当中,分类效果并不明显,没办法更加准确的得到预测效果;同时在分类模型中,数据较少的原因,利用 *K-means* 聚类时,无法准确的找到划分界限,界限相对模糊,但是即使数据较少,依旧可以得到较为理想的分类结果。所以不难推断,在数据较为充分的条件下,可以得到更为出色的亚分类结论。因此,在数据的获取方面,需要更多的投入研究,以支撑算法的推进验证。所以充分的数据积累,在算法的应用当中必不可少,数据越多,越能得到更贴切的统计规律与数值结论。

参考文献:

- [1] 周良知. 影响硅酸盐玻璃风化的主要因素[J]. 大连轻工业学院学报, 1984(1): 34-44.
ZHOU Liangzhi. J Dalian Institute Light Ind (in Chinese), 1984(1): 34-44.
- [2] 王承遇, 陶瑛, 陈敏, 等. 钠钙铝镁硅酸盐玻璃和碱铅硅酸盐玻璃的风化[J]. 硅酸盐通报, 1989(6): 1-9.
WANG Chengyu, TAO Ying, CHEN Min, et al. Bull Chin Ceram Soc (in Chinese), 1989(6): 1-9.
- [3] 赵凤燕, 陈斌, 柴怡, 等. 西安出土若干玻璃器的 pXRF 分析及相关问题探讨[J]. 考古与文物, 2015(4): 111-119.
ZHAO Fengyan, CHEN Bin, CHAI Yi, et al. J Archaeal Cultural Relics (in Chinese), 2015(4): 111-119.
- [4] 冯百龄. 中国出土古代玻璃珠数据库建设与应用[D]. 西安: 西北大学, 2021.
FENG Bailing. Construction and application of ancient glass bead database unearthed in China (in Chinese, dissertation). Xi'an: Northwestern University, 2021.
- [5] 张丽艳, 李洪, 陈树彬, 等. 玻璃的成分和性质的模拟方法[J]. 硅酸盐学报, 2022, 50(8): 2338-2350.
ZHANG Liyan, LI Hong, CHEN Shubin, et al. J Chin Ceram Soc, 2022, 50(8): 2338-2350.
- [6] 李建刚. 基于深度学习的玻璃缺陷检测[D]. 福州: 福建工程学院, 2021.
LI Jiangang. Glass defect detection based on depth learning (in Chinese, dissertation). Fuzhou: Fujian Institute of Engineering, 2021.
- [7] 谢建新, 宿彦京, 薛德祯, 等. 机器学习在材料研发中的应用[J]. 金属学报, 2021, 57(11): 1343-1361.
XIE Jianxin, SU Yanjing, XUE Dezhen, et al. J Acta Metall Sin (in Chinese), 2021, 57(11): 1343-1361.
- [8] 李旭. 五种决策树算法的比较研究[D]. 大连: 大连理工大学, 2011.
LI Xu. Comparison of five decision tree algorithms (in Chinese, dissertation). Dalian: Dalian University of Technology, 2011.
- [9] 何选森, 何帆, 徐丽, 等. *K-Means* 算法最优聚类数量的确定[J]. 电子科技大学学报, 2022, 51(6): 904-912.
HE Xuansen, HE Fan, XU Li, et al. J Electric Sci Technol Univ (in Chinese), 2022, 51(6): 904-912.
- [10] 曹正凤. 随机森林算法优化研究[D]. 北京: 首都经济贸易大学, 2014.
CAO Zhengfeng. Study on optimization of random forests algorithm (in Chinese, dissertation). Beijing: Capital University of Economics and Business, 2014.
- [11] 刘悦, 邹欣欣, 杨正伟, 等. 材料领域知识嵌入的机器学习[J]. 硅酸盐学报, 2022, 50(3): 863-876.
LIU Yue, ZOU Xinxin, YANG Zhengwei, et al. J Chin Ceram Soc, 2022, 50(3): 863-876.
- [12] 刘霏凝, 石竞琛, 王文杰, 等. 材料科学中机器学习算法的应用综述[J]. 化工新型材料, 2022, 50(9): 42-46, 52.
LIU Feining, SHI Jingchen, WANG Wenjie, et al. J New Chem Mater (in Chinese), 2022, 50(9): 42-46, 52.
- [13] 詹祺源. 机器学习的发展史及应用前景[J]. 科技传播, 2018, 10(21): 138-139.
ZHAN Qiyuan. Sci Technol Commun (in Chinese), 2018, 10(21): 138-139.
- [14] CERIOTTI M. Unsupervised machine learning in atomistic simulations, between predictions and understanding[J]. J Chem Phys, 2019, 150(15): 150901.
- [15] JING L L. TIAN Y L. Self-supervised visual feature learning with deep neural networks: A survey J. IEEE T Pattern Anal, 2021, 43(11): 4037-4058.
- [16] 武杰杰. 随机森林算法的应用与优化方法研究[D]. 江南: 江南大学, 2021.
WU Weijie. Research on Application and Optimization Method of Random Forests Algorithm (in Chinese, dissertation). Jiangnan: Jiangnan University, 2021.
- [17] 孙林, 刘梦含, 徐久成. 基于优化初始聚类中心和轮廓系数的 *K-means* 聚类算法[J]. 模糊系统与数学, 2022, 36(1): 47-65.
SUN Lin, LIU Menghan, XU Jiucheng. J Fuzzy Syst Mathematics (in Chinese), 2022, 36(1): 47-65.
- [18] 郑英鑫. 数据挖掘中基于肘部法则的聚类分析在中小出行路线优化设计的应用[J]. 电子世界, 2017(9): 146.
ZHENG Yingxin. J Electronic World (in Chinese), 2017(9): 146.
- [19] Chinese Society of Industrial and Applied Mathematics[EB/OL]. [2022-09-15]. http://www.mcm.edu.cn/html_cn/node/5267fe3e6a512bec793d71f2b2061497.html.
- [20] 张丽艳, 李洪, 陈树彬, 等. 玻璃的成分和性质的模拟方法[J]. 硅酸盐学报, 2022, 50(8): 2338-2350.
ZHANG Liyan, LI Hong, CHEN Shubin, et al. J Chin Ceram Soc, 2022, 50(8): 2338-2350.
- [21] 张伟锋. 斯皮尔曼简捷相关系数与基尼伽玛相关系数的统计特性分析[D]. 广州: 广东工业大学, 2020.
ZHANG Weifeng. Statistical Analysis of Spearman's footrule and Gini's gamma (in Chinese, dissertation). Guangzhou: Guangdong

- University of Technology, 2020.
- [22] 段志强, 裴小龙, 郭庆伟, 等. 多模态混合输入模拟实验过程实现新型 Al-Si-Mg 系合金设计[J]. 物理学报, 2022, 11(18): 1–24.
DUAN Zhiqiang, PEI Xiaolong, GUO Qingwei, et al. J Phys (in Chinese), 2022, 11(18): 1–24.
- [23] 方毅, 陈煜之. 机器学习与线性模型在资产收益率预测中的比较[J]. 统计与决策, 2022, 38(16): 180–183.
FANG Yi, CHEN Yuzhi. J Statist Decision-making (in Chinese), 2022, 38(16): 180–183.
- [24] 宋述芳, 何入洋. 基于随机森林的重要性测度指标体系[J]. 国防科技大学学报, 2021, 43(2): 25–32.
SONG Shufang, HE Ruyang. J Natl Univ Defense Technol (in Chinese), 2021, 43(2): 25–32.
- [25] 曹桃云. 基于随机森林的变量重要性研究[J]. 统计与决策, 2022, 38(4): 60–63.
CAO Taoyun. J Statist Decision (in Chinese), 2022, 38(4): 60–63.
- [26] 李沫. 战国秦汉时期费昂斯制品的制备及铅钡玻璃研究[D]. 北京: 北京化工大学, 2014.
LI Mo. Study on faience simulation and lead-barium glass of the ancient china during the period from warring states to Qin Dynasty and Han Dynasty (in Chinese, dissertation). Beijing: Beijing University of Chemical Technology, 2014.
- [27] 程全中, 谭博仁, 王勇, 等. 机器学习在化学工程中应用研究的若干进展[J]. 化学工程, 2022, 50(9): 23–29.
CHENG Quanzhong, TAN Boren, WANG Yong, et al. Chem Engineer (in Chinese), 2022, 50(9): 23–29.
- [28] 张胜利, 胡扬, 周文瀚, 等. 机器学习在二维材料探索中的应用[J]. 金属功料, 2022, 29(4): 1–21.
ZHANG Shengli, HU Yang, ZHOU Wenhan, et al. J Metal Funct Mater (in Chinese), 2022, 29(4): 1–21.
- [29] 申林, 贾璐阳, 汤典东, 等. 计算化学中的机器学习[J]. 中国科学: 化学, 2022, 52(6): 858–868.
SHEN Lin, JIA Luyang, TANG Diandong, et al. J China Sci: Chemistry (in Chinese), 2022, 52(6): 858–868.