基于机器学习的殷墟花园庄 M54 青铜器 p－XRF 成分数据再思考

黄　希，杨清越，何毓灵

（中国社会科学院 考古研究所，北京 100102）

摘要：随着检测设备的普及与检测精度的提升，文物成分分析数据存量正在急速增加，目前单一目的、单一处理方式的数据研究模式已无法满足精细化考古工作现场对于大规模数据深入分析的要求，基于当下在信息技术领域爆发式发展的机器学习技术，通过机器学习的数据挖掘和模式识别功能，使用现存的安阳殷墟花园庄 M54 铜器 p－XRF 数据进行关联规则挖掘、聚类、预测处理，快速识别出可解释的规律、模式和信息，促进机器学习在考古学研究中的推广与应用。

关键词：机器学习；成分；殷墟；青铜器；聚类；预测

中图分类号：K878　　**文献标识码**：A　　**文章编号**：1001－0238(2023)03－0076－09

DOI:10.16140/j.cnki.ydxk.2023.03.007

引言

安阳殷墟花园庄东地 M54[①]位于安阳殷墟宫殿宗庙区内，整体保存完好未经盗扰，年代属殷墟文化二期偏晚阶段，其绝对年代相当于祖庚、祖甲时期，随葬品精美丰富，对研究殷墟时期的墓葬制度、军队体制、手工业发展等问题具有重要意义。

M54 出土青铜器 200 余件，包括礼器、兵器和生产工具等，形制独特，纹饰精美，制作工艺高超，被誉为中国青铜器制作技术的巅峰之作之一。殷墟青铜器的发现对于研究中国古代社会和文明、青铜器制作技术和思想文化等方面都有着重要的意义。刘煜等[②]对 M54 出土青铜器化学成分及金相组织结构等进行研究，主要采用高锡的铜锡二元合金，兵器锡含量低于礼器，大部分器物铅含量低于 2%，为原料中杂质带入，少部分纹饰精细、器型复杂的礼器（如牛尊、方尊等）铅含量较高，体现出工匠对合金配比与性能之间的关系已经有了充分的认识。

对于青铜器制作技术及腐蚀问题的研究涉及到大量成分数据信息，随着检测设备的普及与检

[**收稿日期**]2023－04－12

[**基金项目**]本文为中国社会科学院青年启动项目（项目编号：2022YQNQD012）阶段性成果。

[**作者简介**]黄希（1993－），女，博士，中国社会科学院考古研究所文化遗产保护研究中心助理研究员，主要从事文物保护科学研究；杨清越（1981－），女，博士，中国社会科学院考古研究所考古大数据资料中心助理研究员，主要从事考古大数据和考古人工智能研究；何毓灵（1972－），男，博士，中国社会科学院考古研究所夏商周考古研究室研究员，主要从事夏商周时期考古发掘和研究。

① 中国社会科学院考古研究所：《安阳殷墟花园庄东地商代墓葬》，科学出版社，2007 年，第 227 页、第 251－252 页。

② 刘煜、何毓灵、徐广德：《M54 及 M60 出土青铜器的成分分析》，《安阳殷墟花园庄东地商代墓葬》，科学出版社，2007 年，第 289－296 页；刘煜、贾莹、成小林、姚青芳：《M54 出土青铜器的金相分析》，《安阳殷墟花园庄东地商代墓葬》，科学出版社，2007 年，第 297－301 页。

测精度的提升,成分分析检测的操作门槛已经大幅降低,越来越多的考古单位配备了以便携式X射线荧光能谱仪(p-XRF)等为代表的便携式分析检测设备,在考古一线即可获取大量的成分分析数据。以往单一目的、单一处理方式的数据研究模式已无法满足精细化考古工作现场对于大规模数据深入分析的要求,目前在处理、识别并分析这些数据背后代表的文物信息、历史信息方面还存在很大欠缺,存在成分分析数据存量急速增加但研究利用率低的问题。基于当下在信息技术领域爆发式发展的机器学习技术,通过机器学习的数据挖掘和模式识别功能,使用现存数据进行聚类、分类、关联规则挖掘等识别出有用的模式和信息,或根据已知特征和模式来识别新的实例数据,对于在考古一线根据大规模的分析监测数据快速识别、量化、区分文物本体材料特征以及可能存在腐蚀病害具有重要的应用价值。

一、机器学习

1.1 机器学习概述

机器学习是一种人工智能技术,通过使用算法和统计模型让计算机模拟人类的学习行为,在不进行明确编程的情况下自动识别和理解已知数据的模式,从数据中自动学习并提高算法性能,找到规律并用于数据预测、分类、聚类等相关任务①。这种学习过程不需要明确的程序指示,而是借助于大量的数据和统计学方法来确定关系和模式②。机器学习主要包括以下几个个步骤③:数据预处理、模型构建与训练以及模型评估,根据不同的任务选择适合的模型,并通过大量的数据对其进行训练,最终得到一个可以准确预测未知数据的模型。

机器学习可以按照学习方式和使用的数据集分成三类:监督学习、无监督学习、和强化学习。本研究主要用到监督学习和无监督学习两种形式。

监督学习(supervised learning)是一种利用有标记的训练数据来进行模型训练和预测的机器学习方法。在监督学习中,需要将输入数据和对应的输出标记作为训练数据来训练模型,建立输入和输出之间的映射关系,使得模型能够根据输入预测出正确的输出。常见的监督学习算法包括:回归分析、决策树、支持向量机、神经网络和朴素贝叶斯算法等。

无监督学习(unsupervised learning)是一种不需要标记或仅少量标记的训练数据,直接从数据中寻找数据集中的特征和模式的方法。在无监督学习中,只需要将输入数据作为训练数据来训练模型,模型会自动学习输入数据的内在结构和规律,并基于此对未知数据进行预测和分类。常见的无监督学习算法包括:聚类分析、降维分析和关联规则挖掘等。

监督学习、无监督学习各有特点和优势④,在本研究中,主要使用监督学习和无监督学习方法,具体包括监督学习中的支持向量回归算法SVR,以及无监督学习的K均值聚类算法K-means和层次聚类算法。在研究过程中,还尝试使用线性回归(Linear Regression)算法、K最近邻回归(K-Nearest Neighbor Regression,KNN回归)算法、梯度提升决策树(XGBoost)算法、随机森林(Random

① 刘霏凝、石竞琛、王文杰、赵瑞:《材料科学中机器学习算法的应用综述》,《化工新型材料》2022年第9期。

② Ceriotti M:《Atomistic machine learning between predictions and understanding》,《Journal of Chemical Physics》2019年第15期。

③ 刘悦、邹欣欣、杨正伟、施思齐:《材料领域知识嵌入的机器学习》,《硅酸盐学报》2022年第3期;Jing L、Tian Y:《Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey》,《IEEE》2021年第11期;G. Ruschioni、D. Malchiodi、、A. M. Zanaboni、L. Bonizzoni,Supervised:《Supervised learning algorithms as a tool for archaeology: Classification of ceramic samples described by chemical element concentrations》,《Journal of Archaeological Science: Reports》2023年第49期。

④ Meng T、Huang R、Lu Y:《Highly sensitive Terahertz non-destructive testing technology for stone relics deterioration prediction using SVM-based machine learning models》,《Heritage Science》2021年第9期;王祉皓、赵芗澈、智群、郭明:《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》,《硅酸盐学报》2023年第2期。

Forest)[①]算法等方法。

1.2 SVR 支持向量回归算法

支持向量回归算法(Support Vector Regression)[②]是一种基于支持向量机的非线性回归算法,用于解决连续型数据预测问题。SVR 通过选择核函数将原始的高维特征空间映射到低维空间中进行分类和回归,找到一个能够最大化边界(函数间隔)和限制条件之间的平衡点的超平面模型,以建立输入变量(特征)和输出变量之间的映射关系,进而对新样本进行分类和预测。

1.3 K-means K 均值聚类算法

K 均值聚类算法(K-means Clustering)[③]是一种基于距离度量的无监督学习方法,它将数据集中的样本按照相似度进行分组,形成 k 个簇。该算法首先随机指定 k 个中心点,然后计算每个样本与中心点的距离,并将其分配给距离最近的中心点所在的簇。接着,重新计算每个簇的平均值或中心点,并将新的中心点作为该簇的代表。反复迭代上述过程,直到簇内所有样本都与其所在簇的中心点的距离最小。

1.4 HCA 层次聚类算法

层次聚类算法(Hierarchical clustering)[④]是一种基于计算数据点之间的相似度量的无监督学习方法,它将相似性(或距离)作为度量,计算每个样本之间的距离,然后将相似度高的数据点依次进行合并,通过慢慢合并最接近的簇,依次形成一个层次化的划分为不同群集(cluster)的聚类图。相似性度量指衡量数据点间相似程度的方法,包括距离、相似性系数、相关系数等,可基于欧氏距离、曼哈顿距离、切比雪夫距离、皮尔逊相关系数等多种距离或相似性手段实现。

二、研究方法

按照一定比例将样本集随机分为训练集和测试集。对属于训练集的文物的相关数值化特征与 p-XRF 中的成分数据组合起来,构成有标签的样本集用于训练模型。使用传统的文物保护研究工作方法,结合前人研究成果,有针对性地对器物腐蚀成分、器物表面土样等进行工艺、成分的预研究。根据腐蚀产状的预研究结果对 M54 出土铜器的器物类型、腐蚀状态、腐蚀程度、文物本体稳定性等相关特征进行定性定量的判断,对每件文物的相关特征分别进行赋值和标记形成样本集,通过机器学习对成分数据集进行分析,寻找可能存在的数据规律模式,并通过成分数据预测对应样本所属的类型。

2.1 腐蚀产状的预研究

在病害宏观认知的基础上,使用常规的文物保护研究方法,包括光学显微分析(Leica DVM6)、扫描电镜分析(Phenom XLG2)、拉曼分析(Horiba XploRA、Thermo DXRxi)、X 射线衍射分析(Bruker D8 Advance)、离子色谱分析(Thermo ICS-5000)、X 射线成像分析(XXQ-2005 型)等,将文物腐蚀产状和成分、腐蚀产物、保存稳定性等特征相对应,建立对文物特征赋值的标准,对样本涉及的每件文物特征分别进行数值化,形成有标签的样本集,用作训练机器学习模型。

2.2 机器学习数据获取与预处理

本研究数据集主要为铜器的 X 射线便携荧光能谱仪分析结果,使用 Thermo Niton XL3T 便携能谱仪对花园庄 M54 出土的 118 件金属器进行检测分析,检测时每个检测点使用金属模式检测,部分采用金属模式和和矿石铜锌两种模式采样,每个检测点采集 3 次,采集时间为 30s,取 3 次读数的平均值作为一组检测结果数据进行分析,计量单位为质量比%,对于部分未检出的化学成分,使用 0 值进行补全,对于成分比例累加和低于 80% 的数据组予以剔除。最终得到有效数据 280 组,其中金属模式 205 组,矿石铜锌模式 75 组。

① 李欣海:《随机森林模型在分类与回归分析中的应用》,《应用昆虫学报,》2013 年第 4 期;Qianqian H、Wei L、Siran L、Jianli C:《Detecting? copper trihydroxychloride with reflectance spectroscopy and machine learning methods》,《Journal of Cultural Heritage》2023 年总第 59 期;Jones C、Daly N S、Higgitt C:《Neural network-based classification of X-ray fluorescence spectra of artists´pigments: an approach leveraging a synthetic dataset created using the fundamental parameters method》,《Heritage Science》2022 年第 10 期。

② 王定成、方廷健、唐毅等:《支持向量机回归理论与控制的综述》,《模式识别与人工智能》2003 年第 2 期。

③ 陶莹、杨锋、刘洋等:《K 均值聚类算法的研究与优化》,《计算机技术与发展》2018 年第 6 期。

④ 贾瑞玉、李振:《基于最小生成树的层次 K-means 聚类算法》,《微电子学与计算机》2016 年第 3 期。

2.3 建立机器学习模型

首先明确本次机器学习的目标是对成分数据集进行分类，寻找可能存在的数据规律模式，并通过成分数据预测对应样本所属的类型，包括预测器型种类、腐蚀程度及有害性等。

基于以上目的，在前期腐蚀预研究发基础上，对样本集的p－XRF数据对应检测位置的各个特征与数据组分别进行标记形成训练集，各特征赋值量化标准如下所示：

a. 文物类别：容器、兵器、杂器；

b. 检测点腐蚀程度：1致密锈蚀、2疏松锈蚀、3点腐蚀、4有害锈；

c. 检测点腐蚀形态：A鼓起的瘤状物、B平整片状锈蚀、C粉末状锈蚀、D有害锈、E装饰绿松石、F修复粘接处、G修复焊接处；

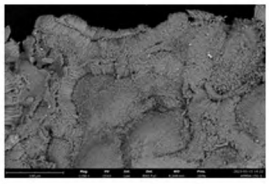
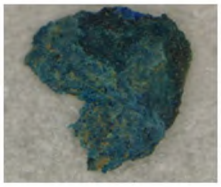
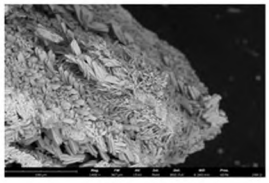
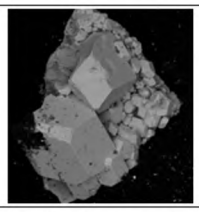
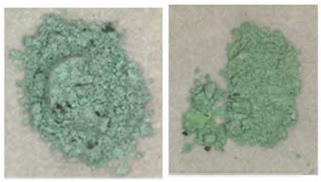
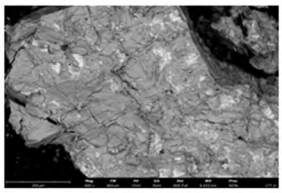
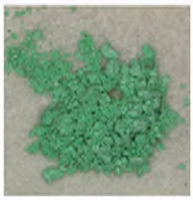
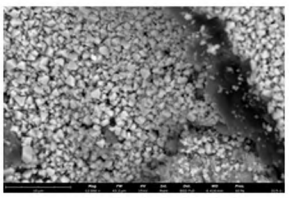
从金属模式中一共有118件铜器，其中兵器35件，容器43件，杂器40件。随机取42组数据为测试集，其余数据组为训练集，训练集包括兵器30件，容器30件，杂器35件，进行机器学习训练。由于矿石铜锌模式数据组总数较少，全部用于作分类算法的训练集。

三、结果与讨论

3.1 腐蚀产状的预研究

在利用文物保护研究技术和手段，对花园庄M54青铜器表面的典型腐蚀产物及分布形态进行检测分析，建立文物成分特征与腐蚀产状的判断标准，为后续对训练集数据特征进行量化赋值步骤建立标准。

表1　M54铜器典型腐蚀产状及成分分析

腐蚀产物显微照片	腐蚀产状	扫描电镜形貌	化学成分
	平整片状锈蚀，小鼓泡，较致密		孔雀石+水胆矾 $Cu_2CO_3(OH)_2$ $Cu_4SO_4(OH)_6\cdot 2H_2O$
	平整片状锈蚀，致密锈蚀		蓝铜矿 $Cu_3[CO_3]_2(OH)_2$
	紫红色晶体颗粒，较致密		赤铜矿 Cu_2O
	浅绿色、浅蓝绿色粉末		锡石 SnO_2
	浅绿色粉末、瘤状物		氯铜矿/斜氯铜矿 $Cu_2(OH)_3Cl$

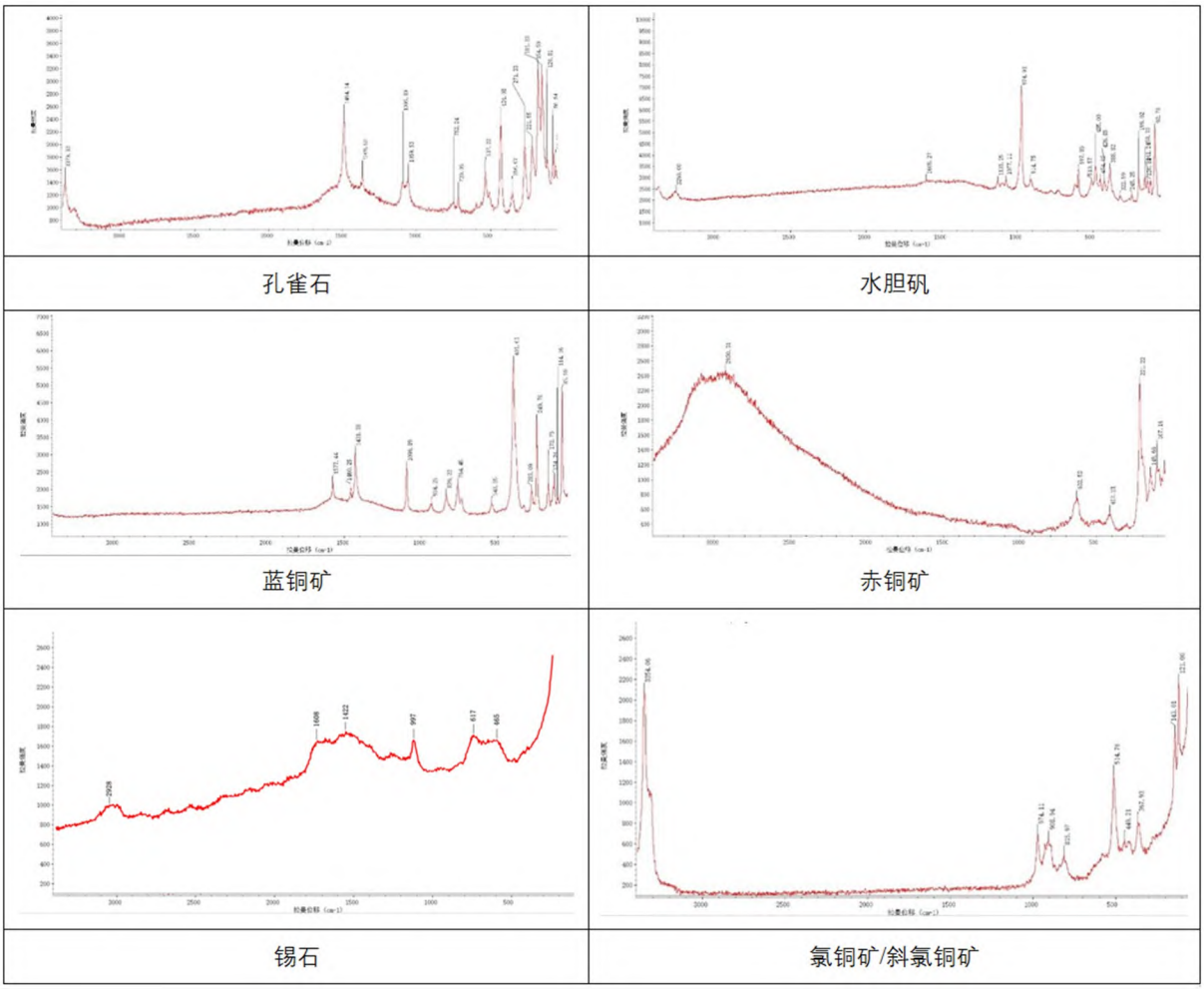


图 1　M54 铜器典型腐蚀产物拉曼光谱分析结果

花园庄 M54 青铜器普遍存在残缺、断裂、裂隙、变形、层状堆积、孔洞、表面硬结物、矿化、点腐蚀等多种病害。表面的锈蚀产物主要有孔雀石、赤铜矿、水胆矾、蓝铜矿、氯铜矿①、锡石等②，其中分孔雀石与氯铜矿存在形式极为接近，都是淡绿色粉末状锈蚀。瘤状物外部及整体大部分为孔雀石，内部靠近青铜基体的部分为赤铜矿。在腐蚀产状调研过程中，发现部分器物出现特殊的腐蚀现象：1、铜铃普遍出现严重的有害锈病害；2、部分铜容器前期焊接位置出现有害锈病害。

3.2　相关性分析

将器型和锈蚀类型这样的分类属性的变量转化成可以量化的变量后，利用 Spearman 相关系数计算法得到数据集中元素与元素、元素与器型、元素与腐蚀之间的相关系数，并导出为热力图，如图 2、图 3 所示。

图 2 为元素 - 元素，元素 - 器类相关性系数热值图，橙色代表正相关，绿色代表负相关，颜色越深代表相关性越强，颜色越浅，代表数值越小。

就元素与元素相关性而言，以 Cu 元素为例，Cu 元素与 Sn 显著负相关（-0.9），可能能够反应出 Sn 原料是人为单独加入，类似的还有 Cu 与 Pb 的负相关关系。微量元素中，Cu 元素与 Au(0.1) 为正相关，Cu 与 As(-0.5)、Zn(-0.5)、Fe(-0.3)、Mn(-0.2) 等均为负相关；Sn 与 As(0.2)、Zn(0.4)、Fe(0.2)、Mn(0.1) 等均为正相关；Pb 与 As(0.2)、Zn(0.2) 为正相关，Pb 与 Fe、Mn 相

① 成小林、杨琴：《五种含氯铜合金腐蚀产物的拉曼光谱及扫描电镜的分析研究》，《文物保护与考古科学》2018 年第 4 期。

② 刘薇、李玲、卫扬波、陈建立：《湖北叶家山墓地出土青铜器的锈层结构研究》，《江汉考古》2019 年第 4 期。主要有孔雀石、蓝铜矿、赤铜矿，部分赤铜矿结晶程度较好，呈现大颗粒紫色立方晶状态，锡石、部

关性为 0。相关性系数 0.1 上下的浮动可能并不绝对,但是相关性系数正负性的差别可以视为明显的区分,铜锡铅三个主量元素与微量元素相关性系数的正负性差别可能与这些微量元素的不同矿石来源有关。

就元素与器型关系而言。就元素与器型相关性而言,除了符合前期研究中已经发现的 Sn 元素在容器(相关性 0.2)、兵器(相关性 －0.1)中的成分差异,Ni、Co、Fe、Cr 元素也表现出分别与容器、兵器正负相关的特点,结合可能引入 Ni、Co、Fe、Cr 等微量元素的矿石来源考虑,古代工匠在不同功能器物冶炼过程中可能存在更多元的合金配比调整。

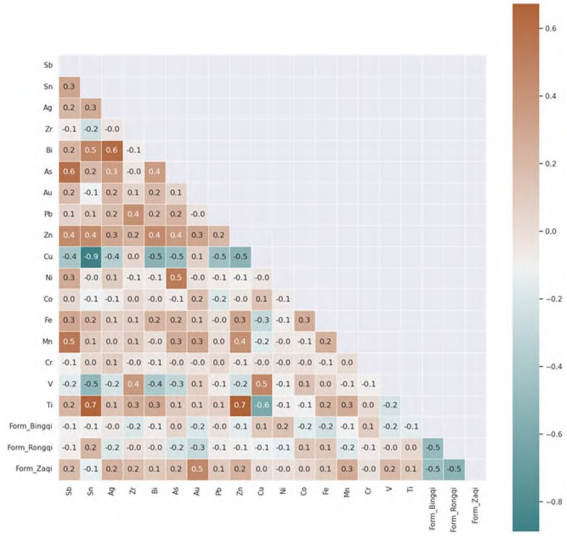


图 2　元素－元素,元素－器类相关性系数热值图

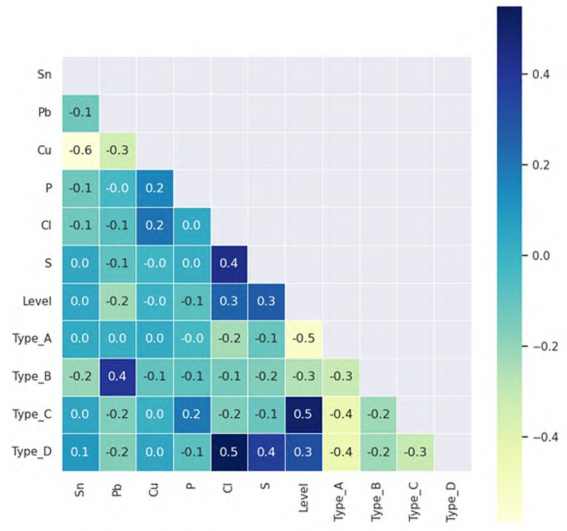


图 3　元素－腐蚀状态相关性系数热值图

图 3 为元素－腐蚀状态相关性系数热值图,深蓝色代表正相关,黄色代表负相关。Level 为腐蚀程度,赋值由 1－4 逐渐加重,腐蚀形态中 E 装饰绿松石、F 修复粘接处、G 修复焊接处为特殊形态,不放入考虑与元素关系。从相关性系数图中可以看出,腐蚀严重程度(Level)与 Cl、S 元素相关性较大(0.3),C、D 腐蚀类型中 A 鼓起的瘤状物、B 平整片状锈蚀、C 粉末状锈蚀、D 有害锈、E 装饰绿松石,B 平整片状锈蚀与 Pb 元素相关性较大(0.4),D 有害锈与 Cl(0.5)、S 元素(0.4)相关性极大。

3.3 模型预测结果

在相关性分析的基础上,分别使用 SVR、Linear Regression、KNN 回归算法、XGBoost、Random Forest 等算法训练模型,以金属模式元素成分与器型关系为训练集,对测试集中 42 组元素数据的器型进行预测判断,并研究模型的精确度。

使用混淆矩阵①(Confusion Matrix)来评估模型的性能。比较预测值和真实值之间存在的差异程度。混淆矩阵通常包含四个条目,分别为真正例(True Positive,TP)、假正例(False Positive,FP)、真反例(True Negative,TN)和假反例(False Negative,FN)。

在本研究中,混淆矩阵如表 2 所示:

表 2　混淆矩阵

	实际为负例 N	实际为正例 P
预测为负例	TN(真反例)	FP(假正例)
预测为正例	FN(假反例)	TP(真正例)

其中,TP 和 TN 为预测情况与实际情况相符的正确预测。

预测结果评估图如图 4 所示,预测目标为“判断元素数据组是否为容器,是为 1,否为 0”,输出结果中,横轴显示为数据组真实情况,纵轴为模型预测结果,深色块表示模型预测与现实结果相符,视为准确预测。

① 于营、杨婷婷、杨博雄:《混淆矩阵分类性能评价及 Python 实现》,《现代计算机》2021 年第 20 期。

图 4　成分 – 器型预测结果

以 SVR 预测结果为例，在现实情况为非容器的对应数据组中（横坐标为 0），有 17 组预测准确，7 组预测错误；在现实情况为容器的对应数据组中（横坐标为 1），有 13 组预测准确，5 组预测错误。整体预测准确度为 71.43%。在仅有 163 组训练集数据的情况下可以达到 70% 以上的准确度。

3.4 分类模型结果

使用 205 组金属模式成分数据组进行分类模型计算，其中，仅使用元素成分结果，主观赋值部分变量（包括器型、腐蚀程度、腐蚀种类赋值）不纳入训练集中。

对训练集数据组进行分类，分别使用 K – means 和 HCA 聚类算法进行分类。

3.4.1 HCA 聚类

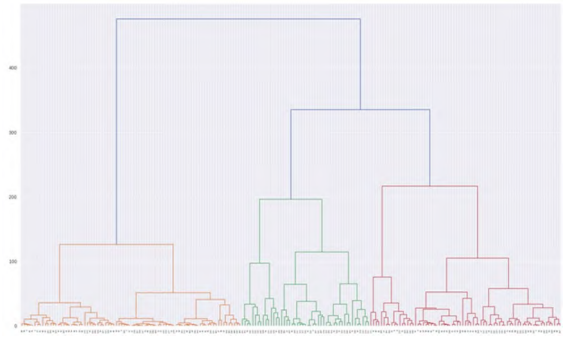


图 5　金属模式成分 HCA 聚类分类结果

HCA 算法以数据相似性为度量，计算每个样本之间的距离，逐渐向上合并形成聚类。分类结果如图所示，如果分成三类，可以按照图中橙色、绿色、红色来区分。其中橙色区域以铜兵器为主（几乎全部铜矛、铜泡、铜刀、铜铃、铜钺），包含部分铜容器（如 2000HDM54：155 铜爵、2000HDM54：84 铜爵等），红色区域以铜容器为主（包括 2000HDM54：392 铜牛尊、2000HDM54：183 方彝，及部分兵器和杂器如 2000HDM54：151 铜刀、2000HDM54：392 铜手等。绿色区域器型特征明显性较弱，容器、杂器和兵器均占有一定比例。

使用 HCA 算法进行聚类分类结果可以大致区分兵器和容器，但特征性较差。

3.4.2 K – means 聚类

K – means 通过将相似的数据点分组为簇来帮助解释数据，利用肘部原则来选择最佳的 K 值，也即是要分成的簇的数量。随着 K 值增加，每种聚类结果的误差平方和（SSE）通常会减小，但是当 K 值达到某个值时，SSE 减少的程度会变得更加缓慢，形成一个“肘部”，选择该肘部所对应的 K 值作为最优聚类数，提高聚类模型的效率与准确性①。由于本次研究的数据组包含多种变量（化学元素种类），先使用主成分分析（PCA）来减少数据维度，并在此基础上运行 K – means 聚类算法。运算结果如图 7 所示。

图 6 肘部原则折线图显示，K = 2 和 K = 3 处出现转折，此后直线斜率明显减小，K = 3 之前为陡降区，K = 3 之后为缓降区，故选择 K = 3，即将数据组分为 3 个类别。如图 7 所示，图中橙色点为三类中理想的典型点，实际并不存在，三个区域可以大致地按照容器、杂器和兵器区分，但三者并不能界限分明的区别开，同时存在较多混杂情况，区分度并不是特别理想。

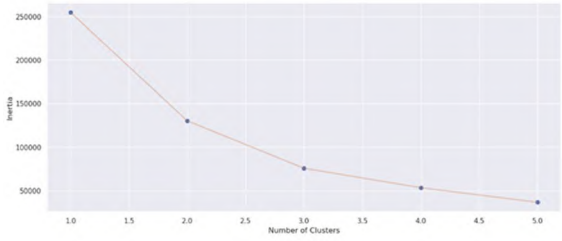


图 6　金属模式训练集肘部原则折线图

① 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》，《模糊系统与数学》2022 年第 1 期。

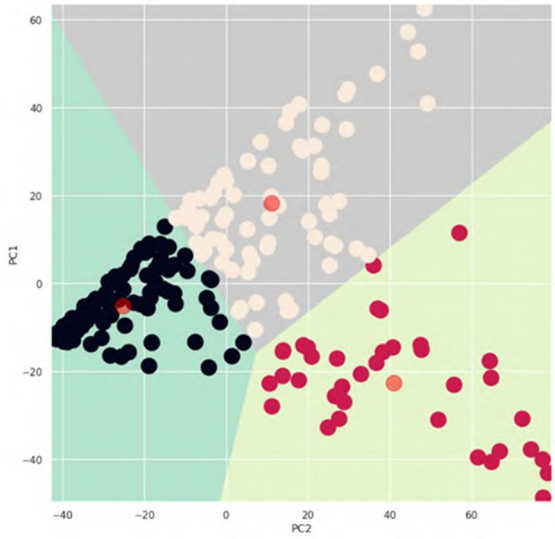


图 7　金属模式训练集分类图

3.5 讨论

3.5.1 样本本身代表性与特异性引入的模型误差

本研究使用铜器表面的 p－XRF 数据进行机器学习的分析，样本本身信息的代表性受到多重影响。首先，由于埋藏过程的长期腐蚀导致表面成分受到腐蚀产物不均匀分布的影响，且文物出土后经过不同程度的清理与不同方式的保护修复处理（粘接/焊接等），以及在博物馆展厅与库房保存环境不同导致的出土后不同的继续腐蚀反应；其次，本研究中将器物分为容器、兵器、杂器三类，分类方式较为粗放，导致同一类别内本身存在较大差异；这些是样本集中事实存在的不同代表性差别。就腐蚀而言，即便是同样的埋藏环境、同样的器型，根据葬仪产生的不同类型的内部盛装物比如肉、骨、酒、植物类等以及殉狗身上的铜铃等，也会产生腐蚀产物的差别，导致成分区别，这是样本本身特异性差异。

在研究过程中还进行了器型－腐蚀状态的相关分析，但是目前结果可解释性较差，根据出土信息可以大概认识到容器与兵器的摆放分别存在一定的集中性，容器集中在椁室南部，兵器集中在椁室的北部和东部，通过对样本信息采集过程的改良与预处理，未来工作可能通过分析显示出显示区域性环境导致的腐蚀病害集中情况。

3.5.2 方法的有效性

在讨论相关性时，考虑整体样本量较少且变量较多，同时存在较多特殊点，使用 Spearman 相关系数而不是传统常用的 Pearson 相关系数。在相关性较强且符合正态分布的情况下，Pearson 相关系数通常是一种有效的计算方式，而 Spearman 相关系数适用于非正态分布和存在外部影响（如离群点）的情况，能够排除一些异常值和极端数据的影响，检测到更广泛的关系类型且不要求数据集满足线性假设。对于受埋藏腐蚀后成分情况更为复杂的出土文物样本成分数据处理，使用 Spearman 相关系数更有助于得出有效结果。

在相关性分析基础上，使用机器学习进行元素－器型预测和聚类分析的过程中，由于目前相对变量而言样本量过少，在建立统计模型时出现了过拟合的问题，在元素－器型关系预测中线性回归、随机森林、KNN、XGBoost 等方法 R2 均大于 1 的现象。这是由于文物样本本身由于腐蚀的发展表面成分与金属基体成分存在一定差异，且本研究采用 p－XRF 检测方法，存在较多的信号噪声可能被误以为是有效信号而造成过度学习。使用 SVR 算法鲁棒性较好，且相对更适用于处理文物成分数据这类高维度数据并减少过拟合现象。

3.5.3 数据改良方式探讨

对于铜器表面成分的影响受到以下几个方面影响：金属基体原始成分、埋藏环境影响、埋藏过程的腐蚀、出土后的腐蚀、修复处理等，在机器学习数据采集过程中需要更明确的问题指向，提高样本采集数量，尽量保证训练集和测试集数据变量均在可讨论范围内。同时在明确研究问题时空范围的前提下，尽量扩大有效训练集本身，如考虑在同一器物上细化采集部位、增加数据采集点位来弥补腐蚀导致的同一器物成分差异，选择腐蚀产状一致的部位进行成分采集与比较；或全面采集同一单位内所有铜器，同时加入层位、区域信息作为新增变量，对训练集中的特征信息赋值更明确；或使用标准统一的更为精确的 p－XRF 量化数据作为训练集，避免采集过程引入的人为误差。

四、结果与展望

机器学习已经广泛应用于数据挖掘和分析、语音和图像识别、自然语言处理、材料学研究等各个领域，应用前景非常广阔。本研究利用机器学

习来研究 M54 出土铜器的 p－XRF 表面成分数据，利用支持向量回归算法 SVR，K 均值聚类算法 K－means 和层次聚类算法分别建立模型，摸索元素－元素、元素－器型之间的数据规律及关系，并使用直观的图形可视化形式进行展示；通过元素－器型预测模型较为准确的预测测试样本元素成分对应的器型特征；使用分类模型对铜器表面成分进行分类，探讨器物成型与腐蚀过程规律，提升了文物成分分析与研究的效率。

4.1 相关性分析

通过计算 Spearman 相关系数讨论铜器表面成分中元素与元素、元素与器型、元素与腐蚀之间的两两相关性。

通过对比 Cu、Sn、Pb 三个主量元素与各微量元素之前的相关性系数，Cu 元素与 Sn 元素的显著负相关（－0.9）及 Sn 元素对容器和兵器的正负相关性差异，与前期研究中围绕锡料的合金配比人为控制的结论相吻合。新发现除 Sn 元素外，Ni、Co、Fe、Cr 等元素也表现出分别与容器、兵器正负相关的特点，可能暗示古代工匠在不同功能器物冶炼过程中存在更多元的合金配比调整模式。在更高维的数据层面体现出不同元素可能的矿料原料来源区别，对于探讨是否存在年代－区域框架下模式化的冶金行为具有一定意义。

另一方面，也证明了使用 p－XRF 对腐蚀铜器进行的无损表面成分数据，在经过合理的数据处理后同样可以体现出铜器的合金配比信息，一定程度上避免了铜器成分分析必须使用金属基体部分而进行的有损取样行为。

4.2 元素－器型预测模型

在相关性分析的基础上，以金属模式元素成分与器型赋值数据为训练集，分别使用 SVR、Linear Regression、KNN 回归算法、XGBoost、Random Forest 等算法训练模型，并使用混淆矩阵来评估模型的准确性。在仅使用 163 组数据作为训练集的情况下，使用 SVR 算法达到的元素－器型预测准确度达到了 71.43%。

本研究中相对变量较多而样本总数较少，模型在训练集上表现良好，但在测试集上表现较差。研究过程中使用的 SVR 方法的 R^2 为 0.971，而线性回归、随机森林、KNN、XGBoost 等方法 R^2 均大于 1，出现过拟合现象。由于出土文物的 p－XRF 数据普遍具有维度高、变量多、回归性较差的特点，使用 SVR 算法可以有效地处理高维度数据并减少存在于传统回归方法中的过拟合（overfitting）现象，在遇到有噪声的数据时也可取得良好的预测结果。由于支持向量是使用部分样本构造出的，所以在精度保证的同时具备一定的可解释性。

建立元素－器型预测模型，可以为破碎严重、叠压复杂的出土文物碎片信息识别和拼对修复提供指导；同时有也有助于更深入了解不同器型的元素特征及背后代表的矿料来源信息。

4.3 分类学习模型

在对成分数据进行分类时，分别使用 HCA 聚类算法和 K－means 聚类算法对 205 组金属模式成分数据组进行分类模型计算，HCA 聚类算法得到多层聚类分布结果，可以大致区分兵器和容器，K－means 聚类算法根据肘部原则确定三相区分结果，但三者之间并不能界限分明的区别开，区分度都不是较为理想。

机器学习在出土文物研究中具有广泛的应用前景，目前，本研究使用的数据量相对变量稍有不足，尚不足与得到理想的模型结果，但在未来工作中，在问题目标明确的情况下更大范围、更精细地获取数据支撑算法的推进验证，预期能够得到更贴近显示的统计规律与预测模型成果。

[责任编辑：郭昱　胡洪琼]