古玻璃风化表面预测内部化学成分方法研究

肖叙昕,张佳怡,谢欣欣,鲁　萍

(西安建筑科技大学理学院,西安 710399)

摘要: 古代玻璃在埋藏过程中的风化会导致其化学成分比例发生变化。针对古玻璃风化表面化学成分还原为内部化学成分数据方法进行研究,构建基于统计学和机器学习方法相结合的风化数据还原模型,依据表面化学成分对样本进行风化程度定义,设计无监督特征选择集成决策方法进行重要特征选取,对重要特征使用 K-means 模型划分风化程度类别,针对风化程度类别的统计特征及成分相关性特点建立回归、映射等风化还原模型,使用该模型还原的内部化学成分满足有效性要求,结果合理性检验使用预测误差和类别还原准确率两个指标,化学成分预测平均准确率约为 67.3%,类别还原准确率约为 90%。

关键词: 无监督学习;特征选择;机器学习;K-means 模型;科技考古;风化还原

中图分类号: TP399　　**文献标志码**: A　　**文章编号**: 1674-8646(2024)16-0058-05

Study on the Prediction Method of the Internal Chemical Composition of Weathering Surface of Ancient Glass

Xiao Xuxin, Zhang Jiayi, Xie Xinxin, Lu Pin

(School of Sciences, Xi'an University of Architecture and Technology, Xi'an 710399, China)

Abstract: Weathering of ancient glass during burial can cause changes in its chemical composition ratio. The study researches the method of reducing the weathering surface chemical composition of ancient glass to internal chemical composition data, and constructs a weathering data reduction model based on the combination of statistics and machine learning methods. Then the study defines weathering degree of samples according to the surface chemical composition, and designs an unsupervised integrated decision method for feature selection to select important features. For important features, the K-means model is used to classify weathering degree categories. Regression, mapping and other weathering reduction models are established according to the statistical and component correlation characteristics of weathering degree categories. The internal chemical components reduced by using this model meet the effectiveness requirements. Two indexes of prediction error and class reduction accuracy are used in result rationality test. The average accuracy of chemical composition prediction is about 67.3%, and the accuracy of class reduction is about 90%.

Key words: Unsupervised learning; Feature selection; Machine learning; K-means model; Science and technology archaeology; Weathering reduction

0　引言

基于文物检测数据的科技考古在文物保护研究[1-2]、文物制作工艺及来源探究方面均具有重要价值。古代玻璃在长时间埋藏过程中会发生风化和腐蚀,内部元素与环境元素进行大量交换导致其成分比例发生变化,从而影响对其类别的正确判断[3]。内部化学成分能更加准确地反映制作技术,但采集内部数据可能会破坏玻璃器物的完整性,因此研究依据表面成分数据推测内部成分数据方法具有重要的应用意义。

目前,关于古代玻璃风化的研究多集中在研究环境条件对玻璃风化的影响[4-5]、可能的腐蚀机理[3]。文献[6-8]基于 2022 年数学建模国赛 C 题的古玻璃成分数据,使用统计学和机器学习方法研究了一批古玻璃表面数据风化和未风化的成分特点及风化还原模型,该数据集均为古玻璃样本的表面数据。从国内外研究来看,少有关于古玻璃内部数据和表面数据化学成分变化规律的统计学研究。

收稿日期:2024-04-07

基金项目:国家重点研发计划重点专项(2019YFC1520200);陕西省大学生创新创业训练计划项目(S202310703)

作者简介:肖叙昕(2003-),女,本科生。研究方向:数据科学与大数据技术;
张佳怡(2002-),女,本科生。研究方向:数据科学与大数据技术;
谢欣欣(2002-),女,本科生。研究方向:数据科学与大数据技术。

通讯作者:鲁　萍(1979-),女,硕士,副教授。研究方向:数据挖掘、推荐系统。E-mail:lping@xauat.edu.cn。

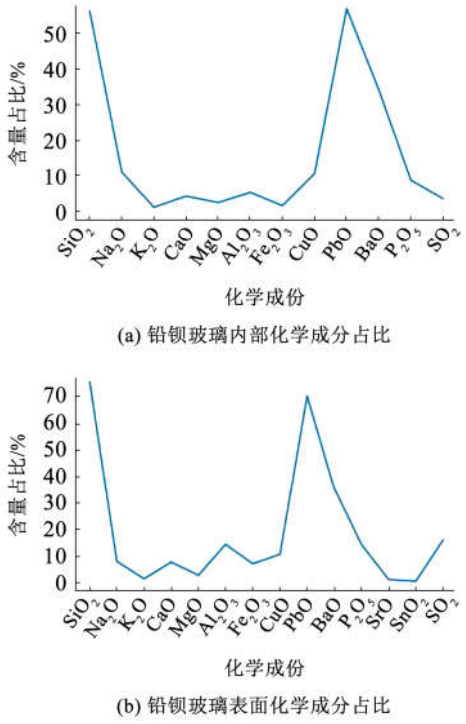
本研究对陕西出土的一批古玻璃表面和内部成分数据进行研究，提出了古玻璃文物的风化化学成分数据的还原方法，依据文物表面化学成分数据预测其内部化学成分数据。

1 相关工作

1.1 数据探索性分析

古玻璃样本包括铅钡类和高钾类，总样本数 84，采集表面数据样本数 74，内部数据样本数 57，对一个古玻璃文物采集其表面数据和内部数据称为匹配数据，总匹配数据样本总数 47，重合化学成分特征 14 个。

以铅钡玻璃为例，考查内部化学成分和表面化学成分统计性指标，总体来看，数据呈现出分散程度大、特征变异系数均高于 15%（均值非 0）、零占比大于 50% 的特征多特点。内部化学成分和表面化学成分占比最大值、最小值、均值如图 1 所示。可以看出，SiO_2 在风化后成分占比最大值、最小值均有较大变化，离散程度更高，PbO、BaO 均值降低，CaO、Al_2O_3、Fe_2O_3 风化后含量占比有所增加。



(a) 铅钡玻璃内部化学成分占比

(b) 铅钡玻璃表面化学成分占比

图 1 铅钡玻璃化学成分占比

Fig. 1 Lead barium glass chemical composition proportion

考查铅钡玻璃匹配数据成分间的关系有以下几种情况：有线性相关性如 PbO，没有相关性如 CaO，有分段映射关系如 Al_2O_3。

1.2 内部化学成分还原方法设计

基于数据集统计描述分析及问题背景分析，玻璃的风化程度不同是其表面化学成分数据离散度较大的主要原因，制作技术不同是其内部化学成分数据离散程度较大的主要原因，因此使用无监督学习方法对样本划分类别后进行分类处理。对表面数据的风化程度进行区分，从不同角度衡量特征价值，进行综合评价及重要特征选取，根据重要特征进行 K-means 聚类，划分类别，确定风化程度标签。对不同风化程度数据集根据其表面、内部匹配数据集的化学成分关系选择风化还原方法，如回归方程拟合、映射函数、分段函数，思路如图 2 所示。

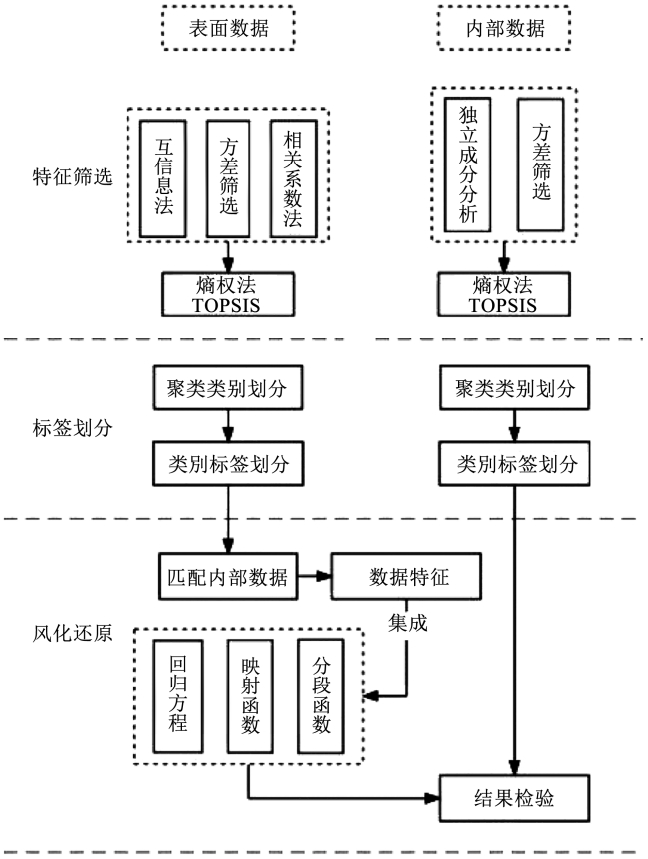


图 2 设计思路

Fig. 2 Design thinking

2 方法描述

2.1 无监督学习特征选择方法

构建一个综合的特征选择框架，结合数据的局部结构特点，尽量降低小样本数据受误差影响大的风险，采取集成决策思路，分别使用方差、互信息、相关系数对特征重要性进行排序[9]，用熵权法 TOPSIS 优劣解距离法进行综合评价，筛选出重要特征。

2.1.1 方差筛选

方差筛选是一种有效的统计指标特征选择方法，方差较大的特征意味着重要性较大。假设数据集中包含 n 个样本和 m 个特征，对每个特征的所有样本计算方差，根据方差大小对特征进行降序排序并绘制折线图，得到特征的重要度指标，以方差图中变化明显的转折点作为阈值，选择阈值前的 k 个具有最大方差的特征作为最终的特征集合。

2.1.2　相关系数法

相关系数是一种衡量两个变量之间线性关系的度量,其值介于 -1 和 1 之间,绝对值越大表示两个变量之间的线性关系越强[10]。计算所有特征之间的 Pearson 相关系数,构成相关性矩阵,为得到每一特征关于其他特征对该特征给出的评价,将得到的相关系数矩阵按行看作新的样本,按列看作特征元素,当某一特征与其他特征的相关性之和越大,意味着该特征在整体数据集的相关性中所占比重越大。因此将每行的相关性之和进行归一化,得到一个反映特征排名重要度的相对指标。根据排名重要度降序排序结果得到特征在数据集中的重要度排序,适用于连续型和离散型数据及分析具有相同尺度的变量之间的相关性。特征排名重要度计算公式如下:

$$rank_k = \frac{\left|\sum_{j=1}^{m} r_{kj}\right|}{\sum_{i=1}^{m}\sum_{j=1}^{m}|r_{ij}|}, k = 1, \cdots, m \tag{1}$$

其中,r_{kj}是第 k 行元素第 j 列的相关系数,r_{ij}是相关系数矩阵。

2.1.3　独立成分分析法

对 n 组样本、m 个指标的数据计算每列的均值和标准差及特征重要度,将计算得到的每一列特征重要度除以数据矩阵的行数,得到特征的重要指标,得到的值越高表示该特征在预测模型中的作用越大,特征重要度计算公式如下:

$$C = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\left|\frac{x_{ij} - \overline{x_j}}{\delta_j}\right|\right)}{mn} \tag{2}$$

其中,$\overline{x_j}$为第 j 列数值的绝对值,δ_j 为第 j 列数值的标准差。

$$\overline{x_j} = \frac{\sum_{k=1}^{n} x_{kj}}{n},$$

$$\delta_j = \sqrt{\sum_{k=1}^{n}\frac{(x_{kj} - \overline{x_j})^2}{n}} \tag{3}$$

2.1.4　综合评价

综合考虑特征的离散程度、互信息及与其他特征的相关性,用熵权法和 TOPSIS 优劣解距离法进行综合评价(图 3)。

通过计算每个特征的信息熵确定特征的权重,信息熵越大表示变量的不确定性越大,权重也相应地越大。依据权重通过 TOPSIS 优劣解距离法计算综合得分,得到最终的特征重要性排序[11]。TOPSIS 综合评价是通过计算每个特征与最优解、最劣解之间的差距,从而对特征进行评价。

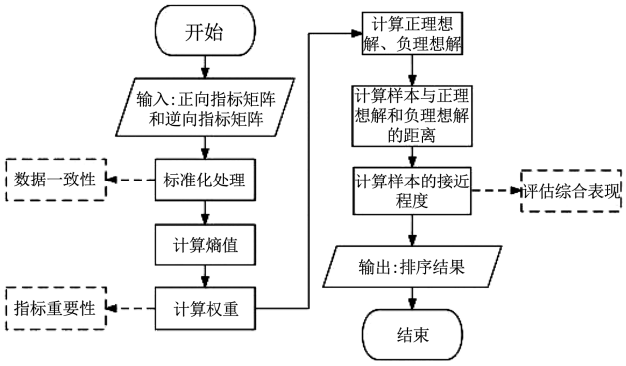


图 3　特征重要性综合评价

Fig. 3　Comprehensive evaluation of feature importance

2.2　无监督聚类方法

聚类算法与特征选择的综合使用能够提升数据分析的准确度和效率,从而对数据进行准确分类。主要分析基于特征排序筛选特征再聚类的方法。

2.2.1　系统聚类

系统聚类是一种无监督学习方法,广泛应用于多维数据的分组,是可以将数据集中的样本分成若干类的聚类方法。其基本思想是从初始状态下将每个样本各自视为一类,根据样本之间的亲疏程度逐步合并距离最近的两个类别,直至所有样本合并为一类。

2.2.2　K - means + +

K - means + + 聚类方法是对传统的 K-means 聚类算法的改进,通过更好地选择初始聚类中心来减少迭代次数并加快算法收敛速度,提高聚类算法的准确性[12]。

2.2.3　聚类结果分析

对聚类结果进行分析主要是为了评估聚类算法的性能和聚类结果的可靠性[13]。簇内平均距离较小,说明聚类结果紧密度较高。如果簇间平均距离较大,说明各簇之间区分明显,聚类效果好。观察轮廓系数的分布,如果大部分簇的轮廓系数接近 1,说明聚类效果较好。

2.3　风化还原方法

成分数据的风化还原是建立表面风化数据与内部数据之间的联系,对呈现一定趋势且回归拟合优度较好的数据采用回归方程拟合方法,根据数据情况对回归方程的次数和元数进行确认。而对于某些数据集中聚合在一个很小的范围内或零占比过高的情况,则适合采用映射方法,考虑到数据存在变异性,采用上、下四分位数作为映射函数的范围。成分数据的特殊性导致数据存在全局分散、局部集中的趋势,应采用分段函数。而在分段函数中根据数据情况对数据进行分段后,需通过回归方程和映射两种方法对分段函数进行构造。如图 4 所示。

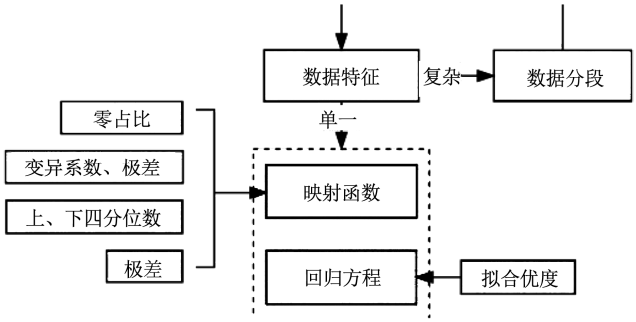


图 4　风化还原方法

Fig. 4　Weathering reduction method

3　实验验证

3.1　表面数据风化程度分析

对表面数据使用方差筛选法、互信息法、相关系数法进行特征重要指标的计算与综合排序，其中在进行方差筛选时需对数据进行标准化。基于表面数据综合评价方法计算特征重要性的结果如表 1 所示。选取 SO_2、P_2O_5、CaO、Al_2O_3、MgO、K_2O、SnO_2、SiO_2、CuO、Fe_2O_3、SrO 作为表面数据聚类的特征依据。

表 1　表面数据综合评价得分

Tab. 1　Comprehensive evaluation score of the surface data

成分	方差法	互信息法	相关系数法	综合评价得分
SiO_2	0.07	0.00	0.03	0.30
Na_2O	0.05	0.03	0.00	0.15
K_2O	0.06	0.01	0.03	0.34
CaO	0.06	0.13	0.04	0.47
MgO	0.06	0.04	0.04	0.37
Al_2O_3	0.05	0.18	0.02	0.38
Fe_2O_3	0.04	0.03	0.03	0.30
CuO	0.05	0.17	0.01	0.32
PbO	0.06	0.00	0.01	0.18
BaO	0.06	0.09	0.01	0.23
P_2O_5	0.08	0.25	0.02	0.59
SrO	0.05	0.00	0.03	0.28
SnO_2	0.07	0.00	0.03	0.33
SO_2	0.04	0.48	0.02	0.64

使用重要特征对表面成分进行聚类，进行 K－means 聚类分析，将样本数据划分为 4 类，对应匹配数据 4 种风化类别及样本数为：1 级 15 个样本、2 级 11 个样本、3 级 5 个样本、4 级 2 个样本。

3.2　风化还原方程的构建

对 4 种风化程度的类别分别建立内部数据还原模型。在每个类别中依据每个特征的数据统计性指标零占比、变异系数、极差等进行综合分析，若特征的内部数据和表面数据有函数关系则直接用回归方程。若数据情况复杂，可分段处理。若成分占比低或 0 占比高则用映射函数。以 2 级风化数据集为例，化学成分零占比如图 5 所示：

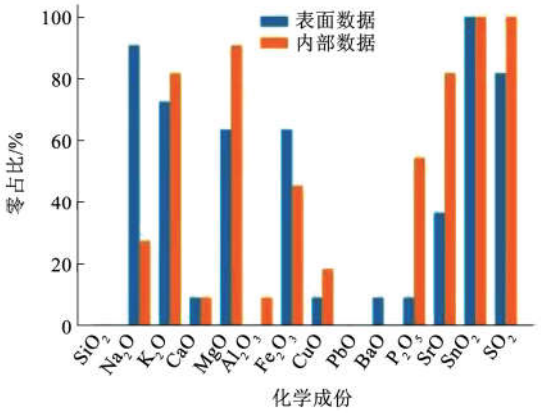


图 5　2 级风化匹配样本数据零占比

Fig. 5　Zero proportion of grade 2 weathering matching sample data

1）回归拟合。若特征的内部数据和表面数据有函数关系则直接用回归方程，BaO 回归方程如公式（4）所示，拟合优度为 0.6127。

$$y = 0.5207x + 7.3515 \tag{4}$$

2）映射。以 Na_2O、K_2O 为例，其特征的描述性统计如表 2 所示。Na_2O 的表面数据接近于 0，零占比高、成分占比低，K_2O 的表面和内部数据含量接近于 0，且极差小，成分占比低，均采用映射还原的方法。映射结果如表 2 所示。

表 2　特征描述性统计指标及还原方法

Tab. 2　Characteristic descriptive statistical index and reduction method

成分	Na_2O		K_2O	
数据类别	表面数据	内部数据	表面数据	内部数据
平均值	0.130	1.70	0.16	0.09
标准差	0.400	1.47	0.29	0.20
极差	1.380	3.82	0.79	0.57
零占比	0.910	0.27	0.73	0.82
成分占比/%	0.195	2.75	0.24	0.15
变异系数	3.160	0.86	1.85	2.15
映射范围	0.47～3.26		0	

3）分段函数。若数据情况较复杂，则分段后分析数据情况。分段后，区间数据有明显单一变化趋势，零占比较高或变异系数大、极差大则构建分段回归和映射函数，SiO_2 分段函数为：

$$y = \begin{cases} -0.0943x^2 + 5.2258x - 37.36, & x \leqslant 33 \\ [35.92, 47.88), & x > 33 \end{cases} \tag{5}$$

2 级类别所有特征方法构建如表 3 所示。

表 3　各特征使用方法

Tab. 3　Use method of each feature

方法	特征
回归拟合	Al_2O_3、CuO、BaO
映射	Na_2O、K_2O、CaO、MgO、SnO_2、SO_2、SrO
分段	SiO_2、Fe_2O_3、PbO、P_2O_5

3.3 风化成分内部数据还原

对风化玻璃进行化学成分还原。对未知样本进行风化程度判别,采用质心法,依据未知样本和4类质心的距离进行判别。使用该类别的风化还原模型进行计算即可得到预测的内部数据。以2级类别进行说明,将匹配数据中的11条风化数据进行带入还原,以样本G-29釉层脱落部位为例,还原结果及真实值对比结果如表4所示,各特征预测值成分和为99.39%,在有效范围内。

表4 匹配样本G-29釉层脱落部位还原结果

Tab.4 Reduction results of the shedding part of the glaze layer in matched sample G-29

特征	预测值	真实值	特征	预测值	真实值
SiO_2	40.69	35.92	CuO	1.88	0.00
Na_2O	0.87	1.17	PbO	54.54	55.27
K_2O	0.00	0.43	BaO	10.59	5.53
CaO	0.94	1.03	P_2O_5	0.39	0.00
MgO	0.00	0.00	SrO	0.00	0.00
Al_2O_3	1.52	0.65	SnO_2	0.00	0.00
Fe_2O_3	0.71	0.00	SO_2	0.00	0.00

3.4 结果合理性分析

对成分内部数据还原的合理性进行检验,采用两种思路:

1)从模型出发进行预测误差分析,即分析各成分预测值和真实值之间的平均相对误差。单个化学成分误差计算公式如下:

$$\varepsilon = \frac{(A - M)}{A} \times 100\% \tag{6}$$

式中,A为真实值,M为预测值。对匹配数据中的表面数据进行风化还原,平均预测误差为32.7%,准确率为67.3%。

2)从应用问题出发进行还原类别准确率分析,依据内部化学成分数据将样本分类,可以理解为每个类别制作技术有所差异,对某个技术类别的风化样本进行内部数据还原后如果仍属于这个类别,则认为还原合理。对铅钡类内部数据选取重要特征:CuO、Fe_2O_3、PbO、Na_2O,聚类后划分为A类和B类两大类,其中A类样本数14,B类样本数17。还原数据类别判别统计结果如表5所示,计算得出正确率为0.903,精确率为1,召回率为0.786,F1值为0.820,该类别分类整体表现良好。

表5 预测值分类类别混淆矩阵

Tab.5 Confusion matrix for classification categories of predicted values

预测值真实值	A类	B类
A类	11	3
B类	0	17

综合来看,预测误差分析准确率为67.3%,说明对单个化学成分的还原预测值和真实值有一定差异,但类别还原正确率为0.903,F1值为0.820,反映出还原预测的整体化学成分的所属类别结构正确。结合问题背景,该风化成分还原模型对解决文物的分类识别问题准确率较高,但对文物复原或仿制来说还需要进一步提高模型的预测准确率。

4 结束语

基于统计学和机器学习构建风化数据还原模型,依据表面化学成分对样本进行风化程度定义,集成多种方法筛选重要特征。对重要特征使用K-means模型划分风化程度类别,在类别基础上通过回归方程拟合、范围映射、分段函数将风化数据还原为内部数据,利用数据误差率及内部数据类别准确率对结果进行检验。结果表明,还原类别准确率较好,可用于分类识别,但预测误差率还有待进一步优化。应深入结合古玻璃专业知识进行分析,降低预测误差率。

参考文献:

[1] 张景科,王玉超,邵明申,等.基于原位无/微损测试方法的砂岩摩崖造像表层风化特征与程度研究[J].西北大学学报(自然科学版),2021,51(03):379-389.

[2] 梁行洲.大足石刻砂岩材料风化程度量化评估[D].兰州:兰州大学,2017.

[3] Dussubieux L, Robertshaw P, Glascock MD. LA-ICP-MS analysis of African glass beads: laboratory inter-comparison with an emphasis on the impact of corrosion on data interpretation[J]. International Journal of Mass Spectrometry,2009,284(1/3):152-161.

[4] Gueli A M,Pasquale S,Tanasi D,et al. Weathering and deterioration of archeological glasses from late Roman Sicily[J]. International Journal of Applied Glass Science,2020,11(01):215-225.

[5] Strugaj G,Herrmann A,Rädlein E. AES and EDX surface analysis of weathered float glass exposed in different environmental conditions[J]. Journal of Non-Crystalline Solids,2021,572:121083.

[6] Zhao J,Tang Z,Deng J. A new method for the composition analysis and classification of ancient glass products before and after weathering[J]. Highlights in Science, Engineering and Technology,2023,40:179-189.

[7] 宛惠,邓明华.古代玻璃制品成分分析与鉴别的统计建模[J].数学建模及其应用,2023,12(02):27-40.

[8] 楼阳,窦雷,卓朝阳,等.古代玻璃成分分析与亚类划分方法研究[J].数学建模及其应用,2023,12(04):73-83.

[9] Solorio-Fernández S, Carrasco-Ochoa J A, Martínez-Trinidad J F. A review of unsupervised feature selection methods [J]. Artificial Intelligence Review,2020,53(02):907-948.

[10] Zhou H, Wang X, Zhu R. Feature selection based on mutual information with correlation coefficient [J]. Applied Intelligence, 2022, 52:1-18.

[11] Chen P. Effects of the entropy weight on TOPSIS[J]. Expert Systems with Applications,2021,168:114186.

[12] Lattanzi S,Sohler C. A better K-means++ algorithm via local search [C]//International Conference on Machine Learning. PMLR,2019: 3662-3671.

[13] Liu T,Yu H, Blair R H. Stability estimation for unsupervised clustering: a review [J]. Wiley Interdisciplinary Reviews: Computational Statistics,2022,14(06):1575.