基于K近邻算法的古代玻璃制品成分分析模型

谷晓蕊　野旭浩　高梓轩　谷　祺　田　志
石家庄铁道大学　河北石家庄　050043

摘　要：本文主要应用t系数测定法、卡方检验、K近邻算法综合性、多角度地对玻璃成分含量进行分析，具有较强普适性[1]。

本文建立基于均值误差修正的K近邻算法预测模型。首先对数据进行预处理，之后采用t系数测定法初步得到表面风化与各指标间的相关程度，再利用卡方检验对其关系进一步分析，得出表面风化与玻璃类型数据的相关关系存在显著性差异，与纹饰、颜色无明显相关性；通过均值法分别得出高钾类与铅钡类风化前后化学含量变化的统计规律；随后引入K近邻算法，得出初步预测结果后，利用均值法所得差值，并引入权重因子进行修正，得出最终预测结果。

关键词：K近邻算法；卡方检验；t系数测定法

An analysis model of ancient glass components based on K nearest neighbor algorithm

Gu Xiaorui, Ye Xuhao, Gao Zixuan, Gu Qi, Tian Zhi
Shijiazhuang Tiedao University, Shijiazhuang 050043, China

Abstract: In this paper, the T-coefficient measurement method, Chi-square test and K-nearest neighbor algorithm are used to comprehensively analyze the composition content of glass from multiple angles, which has strong universality [1].

In this paper, a prediction model of the K nearest neighbor algorithm based on mean error correction is established. Firstly, the data were pretreated in this paper, and then the correlation degree between surface weathering and each index was preliminarily obtained by using the T-coefficient measurement method. Then the Chi-square test was used to further analyze the relationship, and it was concluded that there was a significant difference in the correlation between surface weathering and glass type data, and there was no obvious correlation between the decoration and color. The statistical law of chemical content changes of high potassium and lead barium before and after weathering was obtained by means of the mean value method. Then, the K-nearest neighbor algorithm was introduced to obtain the preliminary prediction results, the difference value obtained by the mean method was used, the weight factor was introduced to correct, and the final prediction results were obtained.

Keywords: K-nearest neighbor algorithm; Chi-square test; T-coefficient method

一、问题重述

分析玻璃文物表面分化与其玻璃类型、纹饰以及颜色之间的关系；根据风化点的检测数据还原其风化前的化学成分含量。

二、问题分析

首先根据题目要求对数据进行预处理。

之后分析玻璃文物表面风化与其玻璃类型、纹饰以及颜色之间的关系，分别通过t系数测定法初步判断它们之间的关系，再进行卡方检验之后，得出不同指标之间的显著性关系。

再根据风化点的监测数据预测其风化前的化学成分含量，简而言之，即为还原已风化样品点的化学成分含量，本文建立基于均值误差下的K-近邻算法模型，利用K-近邻算法，得出初步预测结果后，利用风化前后化学成分含量的均值的差值，结合权重因子对此结果进行修正，最终得出预测结果。

三、模型的建立与求解

1.数据预处理

（1）无效数据剔除

有效数据定义为成分比例累加和介于85%至105%之间的数据即为有效数据。

对不同类型玻璃文物的化学成分含量进行累加，最终发现异常结果，该结果为无效数据、本文直接剔除。

（2）填补缺失值

在对无效数据处理完成后需对空缺值进行填充。在此，本文利用颜色众数对空缺值进行处理填充，将空白处颜色填充为浅蓝色。

2.玻璃文物与指标间的关系

Step1：t系数测定法

t系数测定法，可实现定类指标间相关程度[2]的测量。利用指标变量值的次数来构造测定指标[2]。

Ⅰ.有无风化与玻璃类型的关系

相关系数t_{glass}的计算公式为

$$t_{glass}=\frac{\sum_{j=1}^{n}\sum_{i=1}^{2}\frac{f_{ij}^2}{F_i}-\sum_{i=1}^{2}\frac{m_i^2}{N}}{N-\sum_{i-1}^{2}\frac{m_i^2}{N}} \quad (1)$$

式中，n为指标内种类的个数，N为指标样本的总数，f_{ij}为风化分布的的第i行与指标分布的第j对应的次数，m_i为风化分布第i行对应的总次数。

通过上式可得出，t_{glass}=0.09034，这表明有无风化与玻璃类型相关程度较低。

Ⅱ.有无风化与纹饰的关系

有无风化与纹饰的统计表如下表1所示。

表1　有无风化与纹饰统计表

	A	B	C	总数
风化	11	6	17	34
无风化	11	0	11	22
总数	22	6	28	56

利用上表1与式（1）可得，t_{wen}=0.088235，这表明有无风化与纹饰相关程度较低。

Ⅲ.有无风化与颜色的关系

有无风化与颜色的统计表如下表2所示。

表2　有无风化与颜色统计表

	黑	蓝绿	浅蓝	浅绿	深绿	紫	总数
风化	2	9	16	1	4	2	34
无风化	6	1	8	2	2	3	22
总数	8	10	24	3	6	5	56

同理可得，t_{color}=0.18146，该相关程度相对其他指标较高。

通过上述分析可以初步判断，有无风化与颜色呈现有相对较高的相关性，而其余指标相关性较弱。

Step2：卡方检验

卡方检验可以统计样本的实际观测值与理论值之间的偏离程度，其偏离程度决定了卡方值的大小，卡方值越大，偏离程度越大，反之则越小。

本文利用SPSSPRO进行卡方检验，得出纹饰、玻璃类型、颜色分析结果如下表3所示。

表3　卡方检验结果表

指标	X^2	校正X^2	P
玻璃类型	5.061	3.79	0.024**
纹饰	4.941	4.941	0.085
颜色	8.438	8.438	0.296

注：★代表5%的显著性水平

通过上表3发现，表面风化与玻璃类型其P值为0.085，在5%的置信水平下显著，拒绝原假设，因此其表面风化与玻璃类型数据的相关关系存在显著性差异；表面风化与纹饰、颜色的P值为0.085，0.296，在5%的置信水平下不显著，接受原假设，因此其表面风化与纹饰、颜色数据的相关关系不存在显著性差异。可认为玻璃类型与表面风化有相对较强的关系，其余指标相关性不明显。

表4　卡方检验效应化分析

字段名	ρ	Crammer's V	列联系数
玻璃类型	0.301	0.301	0.288
纹饰	0.297	0.297	0.285
颜色	0.388	0.388	0.362

由上表4分析，得出卡方检验的ρ系数大小，ρ小于0.3即为相关性较弱，根据表3得出，玻璃类型有较为明显的显著性差异，即只需分析玻璃类型的相关系数。ρ_{glass}=0.301相关性较弱，其Cramer’s V值为0.301，玻璃类型与表面风化差异为中等差异。

3.基于均值误差修正的K近邻算法预测模型

本文建立基于均值误差修正的K近邻算法预测模型[3]，首先利用K近邻算法求解出距离某风化点最近的K个同种文物未风化点的平均化学含量，考虑到某些风化点最近同种文物未风化点只有一组数据，所以利用均值误差分析对K近邻算法预测结果进行修正。

（1）指标数据预处理

由于有无风化、纹饰、颜色、玻璃类型均为定类变量，本文将定类变量转为定量变量，对指标数据进行预处理。具体转换为如下式子。

有无风化指标W_{in}可转为

$$Win=\begin{cases}1，文物玻璃有风化\\2，文物玻璃无风化\end{cases} \quad (2)$$

纹饰类型 W_{en} 可转为

$$Wen=\begin{cases}1,\ 纹饰A\\2,\ 纹饰B\\3,\ 纹饰C\end{cases} \tag{3}$$

颜色类型 C_{ol} 可转为

$$Col=\begin{cases}1,\ 蓝绿\\2,\ 浅蓝\\3,\ 紫\\4,\ 深绿\\5,\ 浅蓝\\6,\ 浅绿\\7,\ 黑\\8,\ 绿\end{cases} \tag{4}$$

玻璃类型 T_{yp} 可转为

$$Typ-\begin{cases}1,\ 高钾玻璃\\2,\ 铅钡玻璃\end{cases} \tag{5}$$

（2）模型的建立

Step1：引入K近邻算法思想

K近邻算法，是一种基本的分类与回归方法。采用K近邻算法寻找同种类别中与风化文物距离最近的无风化文物的化学成分进行近似。在风化前数据集中寻找K个最近邻的数据，将K个最近邻的数据求均值，即可预测出风化前化学含量。

首先建立文物信息集合S

S={文物编号，纹饰，玻璃类型，颜色，表面风化} （6）

提取出文物的基本信息 S_{02}={02，1，2，2，2}，寻找出其化学成分含量向量 Che=（SiO_2、Na_2O、K_2O、CaO、MgO、Al_2O_3、Fe_2O_3、CuO、PbO、BaO、P_2O_5、SrO、SnO_2、SO_2）。对基本信息进行变换 S_d={d，1，2，2，1}，寻找同种类型的无风化文物。

转换为集合

$$\begin{cases}S_{20}=\{20,1,2,2,1\}\\S_{45}=\{45,1,2,2,1\}\\S_{46}=\{46,1,2,2,1\}\\S_{47}=\{47,1,2,2,1\}\end{cases} \tag{7}$$

将上述4个无风化文物的化学成分含量集合分别寻找与其相对应的化学成分含量向量 Che_i，之后建立最小距离的目标优化模型。

令最小距离为 $distance$

$$distance_i=\sum_{j=1}^{14}\sqrt{|Che_j-Che_{ij}|^2},\ i=1,2,\cdots,K \tag{8}$$

式中，$distance$ 采用为向量的欧氏距离，Che_{ij} 为第i个同种无风化文物的第j个化学成分含量，令K=3。

建立目标优化模型

$$Dis_{min}=\min\sum_{i=1}^{K}\frac{distance_i}{K} \tag{9}$$

式中，$distance_i$ 为第i个文物编号的距离。

通过上述优化模型，求出前三个最短距离对应的无风化文物。利用所得无风化文物的对应化学成分含量求平均，可得出文物风化前的化学成分含量 $KChe_{02}$。

Step2：均值误差修正

在K近邻算法中，会出现同种类的无风化文物只有一个的情况，所以针对这种情况进行了均值的修正。利用不同种类的均值表，可得出不同种类风化前后均值的差值向量 J_{Ka}、J_{Pb}。利用均值差，修正K近邻算法，在这里只利用均值差修正不为0的数据，减小预测误差。

Step3：引入权重因子

针对上述情况进行权重分配，均值差修正时利用到风化前后的差值进行的修正，很大程度上利用了原有数据，直接进行修正会使预测结果偏差更大。所以本文引入权重因子，对K近邻算法与均值误差修正分别给予合理权重。最终预测为

$$\begin{cases}DChe=\varepsilon KChe+(1-\varepsilon)J_{Ka},\ Typ=1\\DChe=\varepsilon KChe+(1-\varepsilon)J_{Pb},\ Typ=2\end{cases} \tag{10}$$

式中，ε为权重因子，Typ 为玻璃类型，$KChe$ 为K近邻算法得出的最佳结果，$DChe$ 为引入均值差后的结果。这里ε一般取0.5即可。

（3）模型的求解

利用基于均值误差修正的K近邻算法预测模型，利用MATLAB遍历算法，将各风化后的数据进行筛选，最终得出的风化前的预测结果。

根据所得结果，对其进行检验，由于本文考虑成分比例累加和介于85%~105%之间的数据视为有效数据，我们对预测结果进行检验，成分比例累加和均在85%~105%之间，且最小值为85.1%，最大值为101.5%。即最终结果合理。

参考文献：

[1]黄维新.黑钨矿单矿物中化学组份的多元统计分析及其地质意义[J].福州大学学报（自然科学版），1992（03）：123-128.

[2]董西明.两个定类变量间相关系数的计算与分析[J].统计与决策，1997（06）：34-35.

[3]朱冰洁.改进K近邻算法在城市轨道交通客流预测的应用[D].北京交通大学，2019.