智能考古项目报告

南方科技大学·2022年·人工智能导论(CS103)·课程项目

作者：章志轩 12010526、陈志雄 12010302 、宫正 12012803、肖佳辰 12112012、薛丁元 11910213

课程老师：刘江

指导老师：荆志淳、章晓庆

@ 2022.12.18

一、背景

中国作为世界上用玉历史最为悠久的国家之一，在人们的物质生活和精神层面都占据着重要的地位。中国传统意义上的玉是软玉,又名钙角闪石（透闪石-阳起石；$Ca_2(Mg,Fe)_5Si_8O_{22}(OH)_2$)。古人将玉视为君子的化身，是天地的精华或万物的主宰，认为长期佩戴能获得神明的赐福，保佑平安。因此在千年中华文化发展中形成了一种独特的中华民族文化——玉文化。

图1：古玉器

玉文化在中华民族文化中占据显著地位，玉的概念和定义直接影响着我们如何看待和认识古玉的基本特性和其所反映的社会、文化、技术、祭祀、礼仪及宗教信仰等。对古玉的分析将有助于我们理解先民社会形式、文化内涵，揭露古人的思想文化与传承至今的民族精神。比如在众多的矿物岩石（不乏各种各样的宝石和美石）中，中国先民选择软玉来体现重要的文化价值和社会关系，而且这样的选择，持续数千年而不变。一种说法是软玉具有显微交织纤维结构，这种结构赋予软玉高韧性，高致密度，进而决定了软玉特有的温润光泽和质感，被古人视为“玉德”。《说文解字》总结玉有五德仁，义，智，勇，洁，为君子之石。

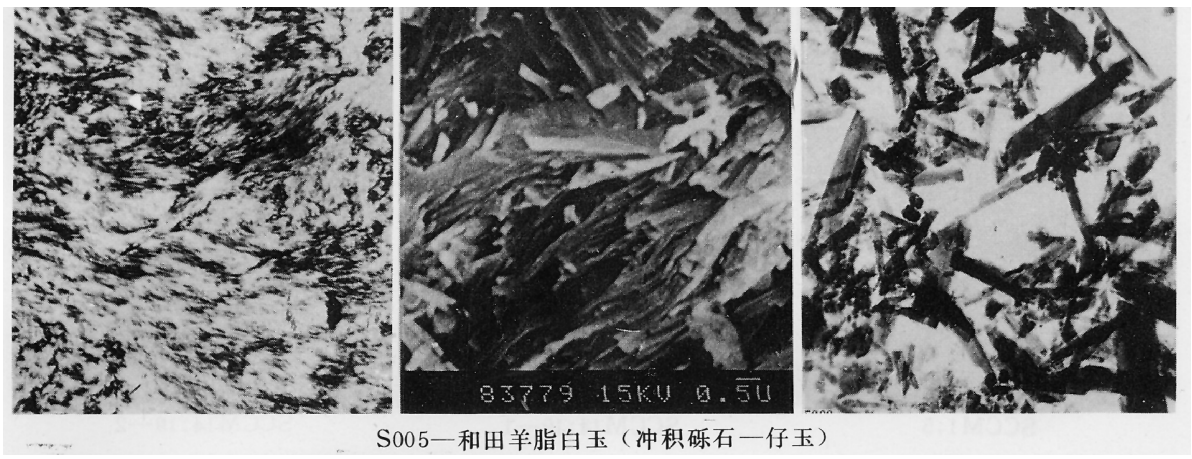

图2:玉的显微结构

二、项目目标

本项目将分析玉石光谱数据，实现人工智能对玉石矿物成分和质地进行分类。

三、实验设计

1. 实验数据

(1) 光谱数据

实验数据是由荆志淳教授提供的安阳花园庄M54墓穴出土的玉器光谱资料，共439个实验对象，光谱数数采取波长1300~2500nm波段，每2nm采取一条数据。同时荆教授还提供了一份去噪数据(去除背景值contiuum-removed)，共231个实验对象，光谱采样与之前相同。

(2) 人工分类数据

荆教授同时提供了一份人工区分的古玉数据，指标为：矿物成分、质地分类、颜色、光泽、半透明度、次生变化和埋葬受沁共7个维度。其中我们只关心***矿物成分***和***质地分类***两类指标。

矿物成分： Np I - 软玉类型 I，Np II - 软玉类型 II，Np III 软玉类型 III，At - 叶蛇纹石，Ct - 绿泥石，It - 伊利石，Dc - 迪开石，Qz - 玉髓、隐晶质石英，UN - 未知矿物或岩石；

质地分类：A - 矿物组成为Np I的器物；B、 B1、 B2 - 矿物组成类同于A, 为Np I, 但含有微量蛭石，B - 未或微弱受沁或次生变化（不包A型 I 式玉管），B1 - 中等或强烈受沁或次生变化（不包A型 I 式玉管），B2 - A型 I 式玉管, 未或微弱受沁或次生变化； C - A型 I 式玉管，矿物组成为Np II；D - 矿物组成为Np III之器物；S - 非玉之矿物或岩石。

2. 实验流程

(1) 数据预处理

首先去除非法数据，然后分别用两种方法独热编码标签，一种是将任务看作多分类任务，对每一种矿物成分都进行独热编码，如图所示：

	At	Ct	Dc	It	Np I	Np II	Np III	Qz	UN
0	0	0	0	0	1	0	0	0	0
1	0	0	0	0	1	0	0	0	0
2	0	0	0	0	1	0	0	0	0
3	0	0	0	0	1	0	0	0	0
4	0	0	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...
420	0	1	0	0	0	0	0	0	0
421	0	0	0	0	0	1	0	0	0
422	0	0	0	0	0	1	0	0	0
423	0	0	0	0	0	0	0	0	1
424	0	0	0	0	0	0	0	0	1

图3:各矿物类型独热码

另一种是将任务看作二分类任务，将矿物成分分为Np类和非Np类进行编码。最后对数据分为训练集和测试集并进行标准化处理。

(2) pca数据压缩：

- 将原始数据按列组成n行m列矩阵X
- 将X的每一行（代表一个属性字段）进行零均值化（去平均值），即减去这一行的均值
- 求出协方差矩阵 C= 1/mX*XT
- 求出协方差矩阵的特征值及对应的特征向量
- 将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P（保留最大的k各特征向量）
- Y=PX 即为降维到K维后的数据

代码流程大致如下：

- 去除平均值
- 计算协方差矩阵
- 计算协方差矩阵的特征值和特征向量
- 将特征值从大到小排序
- 保留最上面的N个特征向量
- 将数据转换到上述N个特征向量构建的新空间中

(3) 模型/算法选定

机器学习上我们选择SVM和随机森林两个算法进行实验，深度学习上我们选择MLP进行实验，原因如下：

- 随机森林：随机森林可以生成多个决策树进行训练，最后结果取各个树的众数。它实现简单，计算量小，可以处理更多特征的数据。随机森林和SVM都是专对物品分类的模型。
  -
  SVM：SVM则是在空间中找到将样本分开的最好的一个超平面，所以它只能对一个标签进行0/1分类，并且样本数量多的时候效率不高，但是它在小样本下表现更好，也无需依赖整个数据，无局部极小值问题。但是在实际使用中，因为Np类型占比过高，导致对此单一分类很难看出区别。权衡之下，我们用了随机森林对全标签进行分类，而SVM只做了矿物成分的Np与非Np的预测。
- MLP：多层感知机是最经典的神经网络。使用中间隐含层、激活函数和反向传播进行对样本的学习，从而实现预测。它的好处是更加普适化，实现简单。我们也用它做了全标签的分类，并与随机森林做了简单比较。

(4) 设计实验

因为需要预测矿物成分和质地分类两个标签，我们将其视为两个任务，分别进行分类预测，再将结果合并成最终结果。然后使用以上三种模型/算法进行实验并对比结果。

(5) 训练

随机森林和SVM的训练过程都是直接调用sklearn库中的函数使用默认参数进行训练，MLP的训练过程则是先定义模型，模型为五层每层100个神经元，然后使用fit函数进行训练。

3. 实验结果

结果的好坏由模型的预测准确率决定。以下是实验结果：

预测准确率	随机森林				MLP				SVM
	矿物成分	质地分类	同时考虑	综合	矿物成分	质地分类	同时考虑	综合	矿物成分
原始数据	0.92	0.84	0.88	0.88	0.96	0.9	0.88	0.93	0.96
去噪数据	0.83	0.75	0.8	0.79	0.82	0.81	0.6	0.81	0.98
原始，且所有Np类视为同一类	1	0.85	0.91	0.92	1	0.91	0.86	0.96	——
去噪，且所有Np类视为同一类	0.88	0.74	0.83	0.81	0.9	0.79	0.75	0.84	——

图4:实验结果

图5:矿物成分分类散点图

1. 左为原始数据的数据分布，右为降噪处理后的数据分布
2. 从上到下，第一行是Np和非Np（Np为蓝）；第二行是Np类合并的7类矿物（Np类为蓝）；第三行是全部9种矿物（Np l 为蓝）
3. 散点图基于 $echarts.js$ 库（一个纯JavaScript的图表库）绘制，由于数据量并不是很大，我们直接将数据筛选分类后拷贝进 $javascript$ 文件中绘制图表。Echarts官方网址： https://echarts.apache.org/zh/index.html，可访问其网址完成对应配置或在线直接使用。

   图表绘制代码见上传文件 `scatter-simple.html`

4. 分析

通过矿物成分分类散点图，我们可以清晰的看出同类矿物有集中分布的趋势，且区别明显。但对于同属于Np大类的（Np I、II、III），我们不能够很好的区分，在原始数据还可以看出Np I(蓝) 和Np II(绿)的区别，但在右边图中已经混合在一起。对于Np类的区分主要出于其它维度的考虑(事实上它们在外观包括颜色、光泽等有较大的区别)，所以在矿物成分维度的区分不明显是正常的，它们的区别在于一些杂质(不属于软玉的部分)有所不同，这带给了玉器不同的物理特征进而被区分。由于这部分不是本项目的重点，所以本文不再过多探讨。

对于模型的分析，首先是SVM模型的实验失败。尽管它的预测准确率很高，但我们分析数据，发现Np类别的实验对象远多于非Np类别，模型只需要一直预测1，就会得到极高的准确率，这显然是不合理的。事实上模型对1的预测全对了对0的预测只有0.92左右。由此，SVM的实验结果并没有特别多的参考意义。

对比随机森林和MLP，可以发现MLP的预测准确率高于随机森林，在矿物成分和质地分类上具有优势。其原因可能是我们使用的数据集太小，或者非Np类的数据量太小，导致随机森林不能产生很好的分类。尽管如此，我们还是可以发现随机森林占优的地方。比如当我们将矿物成分和质地分类同时考虑时，每条训练数据的维度都会增大，这会提高训练难度。自然的类似于MLP，预测结果就发生显著降低(0.81→0.6)，但随机森林基本保持着精度不变，和直接综合的结果等同。

接着是数据集的影响，本实验分了两类数据：原始数据和去噪数据。实验结果显示当使用去噪数据时，预测精度会下降。其最直接的原因是数据量的减小(439→231)，其仅为原始数据的一半。理论上当光谱数据去除背景值后，将减低测量误差(背景值不可能完全去除，误差也不可能为0)，有利于提高分析结果的稳定性和准确度。为此我们重新设计了一个实验：在原始数据中随机挑选231条，重新进行训练并预测。得到的结果为：

预测准确率	随机森林				MLP			
	矿物成分	质地分类	同时考虑	综合	矿物成分	质地分类	同时考虑	综合
原始数据	0.85	0.73	0.81	0.79	0.83	0.78	0.77	0.81
去噪数据	0.83	0.75	0.8	0.79	0.82	0.81	0.6	0.81
原始，且所有Np类视为同一类	0.97	0.7	0.86	0.84	0.97	0.77	0.81	0.87
去噪，且所有Np类视为同一类	0.88	0.74	0.83	0.81	0.9	0.79	0.75	0.84

图6:相同数据量下的实验结果

可以看出，随着数据量的减小，预测准确率确实有所降低，原始数据与去噪数据的预测结果基本等同。但若是视所有Np类为同一类，那么原始数据的结果依旧高于去噪数据。因此前面关于数据量的推测并不完全准确，去噪数据存在其它因素导致精度下降。抛去去噪过程的影响(因为去噪工作并不是我们完成的)，我们重点分析了去噪数据的数据分布，发现去噪数据的Np与非Np的数据量差距更加巨大。对于非Np类，随着负样本增多，精度上升，召回率上升；对于Np类，负样本减少，精度下降，召回率下降。由于数据集以Np为主，当Np的预测精度下降，整体的预测结果也就下降。这也就造成了去噪数据预测结果反而不如原始数据。

所以，最终结果为使用原始数据的精度最高，因为数据量大且正负样本相对更均衡。且若是仅关注Np大类，精度可以接近100%，是十分理想的预测结果。

四、改进方案&总结

使用MLP模型能够得到由于随机森林的预测精度。对于矿物成分的预测，最高的预测精度为100%，质地分类的预测精度为0.91，整体预测精度为0.96。我们选择模型时有考虑到模型的实现难度、训练难度以及是否适合矿物分类三个要素。这给了我们很好的实验空间，代码复用率高。当然这里肯定存在更加适合的网络模型，比如可以将MLP替换成CNN，或是其它适合的网络模型。其次是在数据集的选择上，本次我们使用了现成的数据集，这固然减少了我们的工作量，但也难以得到我们理想的数据样本。

就如本次数据集的正负样本差距过大，造成后续实验的出错，未能达到理想预期。为此，在数据收集以及预处理上下功夫将显著提高我们的实验结果。

本次项目我们只实现了玉石的矿物成分和质地的分类。后续我们将应用本次实验的结果结合考古领域专家对玉和玉料来源进行分析。这将有望得出外来玉器判别进而得出古代社会文化交往的证据等可观的研究结果。再进一步可以研究玉器其它特征属性(颜色、光泽等)，并尝试提取特征波段以发现古人识玉的秘密。最后感谢荆志淳教授和章晓庆博士提供的帮助，在你们的引导下，我们都从本次项目学到了很多知识与经验。