# 基于XXXX的物体指纹图谱的分类与识别

## 摘要

研究目的：

**本文研究已有的处理完成后的物体指纹特征，问题一在通过识别出已给9个物体指纹特征后，提取特征值得到三种类别的分类特征机制。并通过提取余下的5个未分类物体指纹特征，将其归类到以上三大类中从而完成指纹分类。问题二是将给出的14个测试指纹样本划分类别，分析这些物品指纹是否属于不同的大类。**

建立模型思路：

**针对问题一，先从A\B\C三大类中用可视化方式观察得到三种类别大致的自变量——测量值xy走势图。之后对每个样本的数据进行去噪预处理，分析每种趋势中的几何拓扑结构、衍射峰位置、衍射峰形态以及强度等变量，综合后得到相对应的特征值。分别用单特征序列分布相对熵，多特征相似度投票和多阶段特征匹配的概率三种模型建模分析，确定三种模型所得到的三大类分别所包含的指纹图谱的特征向量范围，使用（检验方法）从而得到分类依据。再以此分类依据对待分类的指纹样本进行计算得到特征向量并得到分类结果。**

**问题二分别用层次聚类学习，GCN聚类模型和LDA聚类模型进行聚类分析，得到三种不同的聚类结果后横向对比分类结果，使用问题一中的数据进行检验并确定选择一个效果最佳的聚类方式，从而得到分类依据。**

模型求解和结果：

**在问题一中，本文最终确定使用多阶段特征匹配的概率判别模型，利用（有什么已有的理论知识用到这个模型建立过程中的），使用python工具编码分析，计算出的结果是第10、14分在第三类，第11分在第一类，第12、13分在第二类。在问题二中，本文最终使用LDA聚类学习模型，利用（理论知识同上），通过使用python工具，得到三大类。**

建模特点：

模型优缺点，创新之处，算法特点，模型检验，结果检验，灵敏度分析，稳定性分析等，推广性如何。

**关键词（黑体不加粗小四号）：** **指纹图谱，概率损失，LDA聚类学习**

## 一、问题重述

**问题背景：**

**指纹识别是将识别对象的指纹进行分类比对从而进行判别。指纹识别技术作为生物体特征识别技术之一在新世纪逐渐成熟，进入了人类的生产生活领域。** **每个物品都有自己独特的特征，包括物理状态和化学状态上的不同，使用功能和存在领域的不同，与外界进行物质交换的不同等等都可以作为特征被记录下来。因此，如果能够通过一种特征识别技术得到物体的一些特征后可以快速对其进行识别并判断出物品的种类，可以方便我们了解许多还在探索中的事物，同时也可以加快分类和识别领域的发展。**

**要解决的问题：**

1. **利用已经给出的A/B/C三类物品的指纹图谱，提取三类物品的特征并构造分类方法并检验构造方法的效果。从中选择一种效果最好的方式来完成对其他5个未标明类别指纹图谱的分类。**
2. **使用构造的分类方法对给出的14个其他类指纹图谱进行分类，得到结果。**

## 二、问题分析

模型如何选择：

**问题一中提到通过认识一种物品的指纹图谱来从宏观上对物品进行识别和分类，因此运用已给出的三类已知物品的指纹图谱来构建提取特征的模型，并通过一些方式对这三种特征进行区分。通过观察所给出的自变量和测量值数据在可视化后在xy图像上的表现，发现分类出的三种类别其在走势图上有明显的差异，而相同的类别中的指纹图谱数据在走势和峰值与偏度等特征上则偏向相同。**

**在这个基础上我们选择了三种模型来分析，在前两种模型过渡下确定多阶段特征匹配的概率判别模型作为最终分类模型（接：怎么求解）**

**问题二中需要对14个待分类样本划分类别，因此模型应当是一种无监督的学习聚类模型。考虑到每个样本中需要考虑的特征元素，为了使聚类效果明显，同一种类别数据的投影点尽可能的接近,而不同类别的数据的类别中心之间的距离尽可能的大，我们采用LDA模型来对样本点无监督分类。（接：怎么求解）**

模型怎么求解：

**问题一：对指纹图谱进行分类转化为通过对指纹图谱自变量与测试值之间的关系图形中含有的一些特征值进行计算和区分获得不同类别指纹图谱的不同特征，例如峰值数量特征，峰值位置特征等。相同方式计算出待分类的指纹图谱的对应特征，并通过获取不同类别的多项特征，一一进行比较，最后将待分类样本划分到概率最大的类别中。**

**问题二：将指纹图谱划分类别转化为聚类学习，通过LDA模型得到相对较准确的分类依据，并且通过模型检验确认模型聚类效果良好。**

## 三、模型假设

**模型假设如下：**

**假设一：根据题目要求，假设除了给出的自变量以外，指纹图谱没有其他用来表示特征的自变量。**

**假设二：**

**根据题目要求，假设所有的指纹图谱均不受外界影响。**

**假设的要求：**

1、论文中的假设要以严格、确切的数学语言来表达，使读者不致产生任何曲解。

2、所提出的假设确实是建立数学模型所必需的，与建立模型无关的假设只会扰乱读者的思考。

3、假设应验证其合理性。假设的合理性可以从分析问题过程中得出，例如从问题的性质出发做出合乎常识的假设，或者由观察所给数据的图象，得到变量的函数形式，也可以参考其他资料由类推得到。对于后者应指出参考文献的相关内容。

**好的假设：**通过文献和互联网查到的数据真实可信。

**差的假设：**不合理假设、没必要假设、需要检验的不定论假设、事实假设。

## 四、符号说明

## 五、模型建立

模型建立写作要求：对于每一个模型的建立，需要写出的内容包括：问题分析→公式推导→基本模型→最终或简化模型。

1、基本模型：要有数学模型、数学公式、方案等，要求完整，正确，简明。

2、简化模型：要明确说明简化思想和依据，简化后的模型尽可能给出。

加分项（能在规定时间内做完后还有足够时间的再考虑加分项）：

2、对于同一问题使用两个或以上合理模型进行求解。避免出现单纯罗列模型，又不做对比和评价的现象。

**5.1 数据预处理**

**5.1.1 问题一数据预处理：**

**首先是对数据缺失值、异常值的分析和修复，第一问的三个第B类取样数据的特征维度比其他样本都多了一倍（4500个特征点），由于其B类在那个区域的数据特征值方差较小，且同属一类对分类和识别无贡献，我们选择将其删除，所有的样本统一为2250个特征点。**

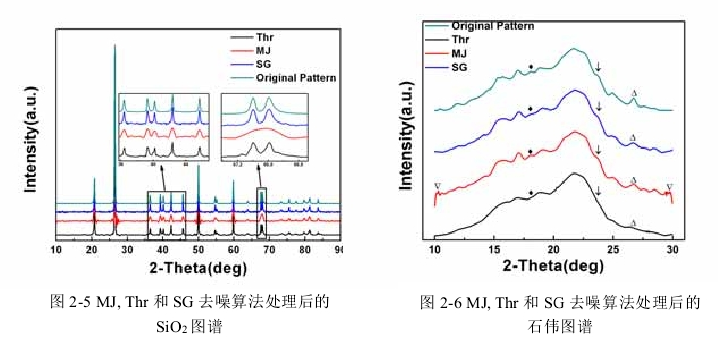
**对数据进行一定的可视化观察后，结合物品指纹图谱技术，我们发现所提供待分类样本的特征数据和X射线衍射图谱数据极为类似。在这种假设前提下，那么我们的数据自变量为衍射角度，因变量则为衍射强度，且结合原始衍射数据曲线的几何拓扑和强度范围，可以知道待分类的样本都是非晶态[]。**

**基于这些先验知识，我们按照X射线衍射图谱预处理的一些传统方式，对数据进行了一定的预处理：平滑处理、扣除背景、图谱寻峰。**

**5.1.1.1 平滑处理**

**我们采用了Savitzky-Golay最小二乘滤波器在角度域内对图谱进行平滑处理，Savitzky-Gola最小二乘滤波器是一种数字滤波器，用于对一组数据点进行平滑处理，能够在不造成大的信号失真情况下提高信噪比值。**

**在去噪之后，我们图谱数据的“毛刺”变少了，整体的曲线变得较为平滑。**

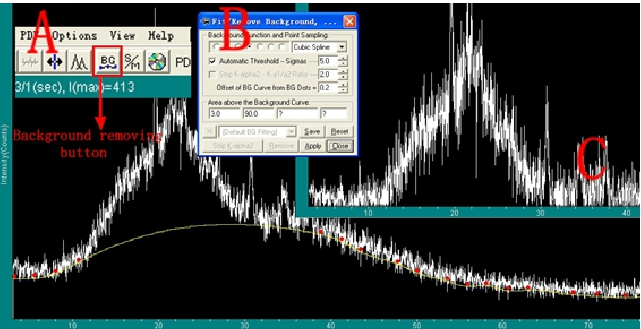


**5.1.1.2 背景扣除**

**背景扣除，也称基线扣除。进行去噪处理后的XRD图谱背景线仍能看到一些小的数值起伏，这些起伏不利于图谱背景的扣除和弱峰的辨认。在进行某些处理前，必须要作背景扣除。**

**基于小波分析，我们对图谱进行背景扣除的处理。**

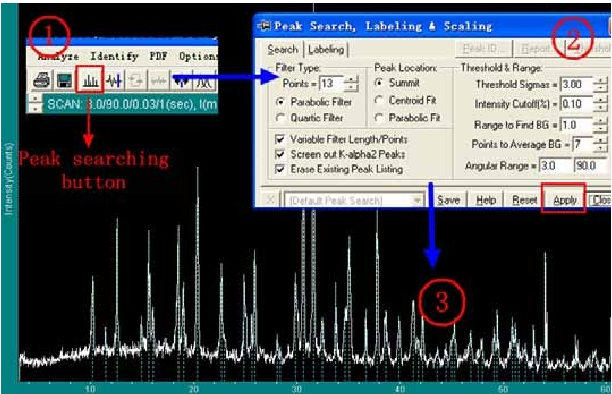
**# 前后对比图**



**5.1.1.3 图谱寻峰**

**衍射峰，是X射线衍射（XRD）技术中非常重要的概念，其能够反映物质的一些特征，帮助我们进行鉴别和分类。在传统的物相分析中，有许多常用的寻峰算法：对称零面积卷积法、简单比较寻峰法、二阶导数寻峰法、协方差法寻峰、线性拟合寻峰法。本文采用线性拟合寻峰法进行寻峰，原因是其简单易用且效果让人满意。**

**# 寻峰的结果**



**5.1.2 问题二数据预处理**

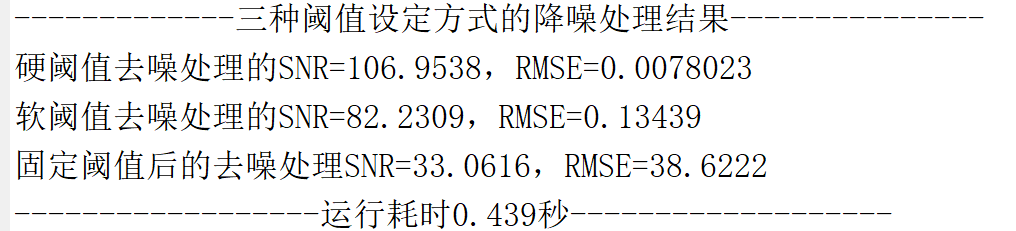
**首先使用代码对数据进行小噪去波，分别采用固定阈值，硬阈值，软阈值三种方法进行处理，结果如下，并分别求的每种去噪方法的信噪比（SNR）和均方根误差（RMSE）。**

**判断标准为信噪比越大，说明混在数据里的噪声越小，均方根误差越小，处理后的数据越接近原数据，更加准确。**

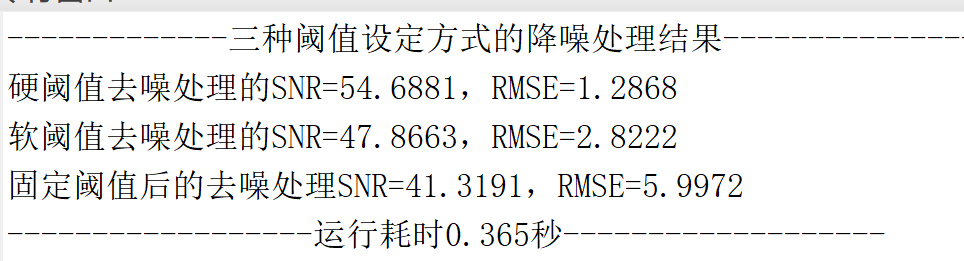
**对前四个数据进行判断，结果如下：**



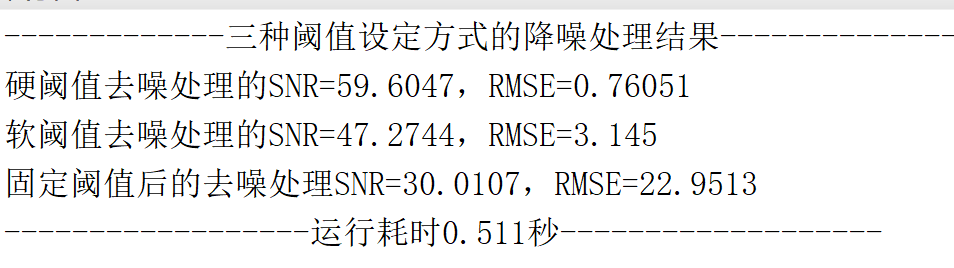
**图5.1.2(1)**



**图5.1.2(2)**



**图5.1.2(3)**



**图5.1.2(4)**

**显然采用硬阈值小波去噪方法明显优于其余两种算法，故在接下来的数据处理中只采用硬阈值处理后的数据。**

**此外分析中发现test4的数据中共用3750项，比其他的2250多了正好2/3,于是考虑对其拉伸。具体方法为将数据以相邻五个为一组，合计750组；接着取每组的前三个数的平均值，中间三个数平均值，后三个数的平均值，合计得到750\*3=2250个数。**

**5.2 问题一的模型建立与求解**

问题分析→公式推导→基本模型→最终或简化模型

**5.2.1 问题分析**

**5.2.2 理论基础**

（1）相对熵

（2）数值微分

**5.2.3模型建立**

针对第一问的有监督聚类问题，我们提出了三种各具特色的模型方案。

**5.2.3.1 基于单特征序列分布相对熵的分类判别模型**

**5.2.3.2 基于多特征相似度投票的分类判别模型**

**5.2.3.3 基于多阶段特征匹配的概率判别模型**

**5.2.4 模型求解**

**5.3 问题二的模型建立与求解**

问题分析→公式推导→基本模型→最终或简化模型

**5.3.1 问题分析**

**5.3.2 理论基础**

（1）层次聚类

（2）GCN

（3）LDA

**5.3.3 模型建立**

针对第二问的无监督学习问题，我们建立了基于层次聚类、GCN、LDA的三种无监督聚类模型。

**5.3.3.1 基于层次聚类的聚类模型**

**5.3.3.2 基于GCN的聚类模型**

**5.3.3.3 基于LDA的聚类模型**

**5.3.4 模型求解**

## 六、模型检验

1. 类中曲线的代表性实验测试

2. GCN的阈值参数

3.截断峰参数K

## 七、模型评价

各方案的对比与小结

## 八、模型推广

## 九、参考文献

## 十、附录