# 基于混合分类算法的物体指纹图谱的分类与识别

## **摘要**

**关键词：**

111s

## 一、问题重述

## 二、问题分析

## 三、模型假设

## 四、符号说明

## 五、模型建立与求解

**5.1 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，对数据进行了平滑处理。

**5.1.1 平滑处理**

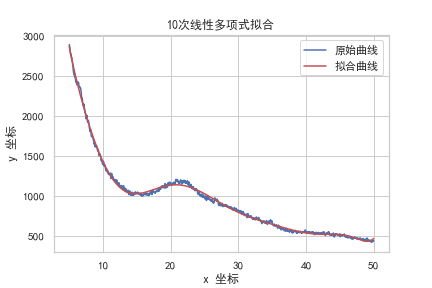
我们采用了Savitzky-Golay最小二乘滤波器在角度域内对图谱进行平滑处理，Savitzky-Gola最小二乘滤波器是一种数字滤波器，用于对一组数据点进行平滑处理，能够在不造成大的信号失真情况下提高信噪比值。

经过Savitzky-Golay最小二乘滤波器处理后，图谱数据的“毛刺”变少了，整体的曲线变得较为平滑。

|  |  |
| --- | --- |
|  |  |

**5.1.2 最终结果**

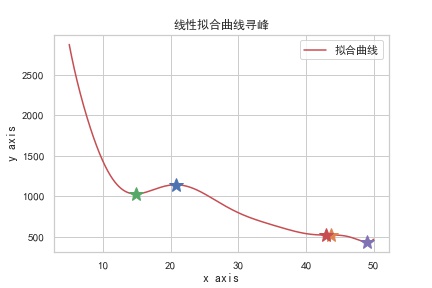
在经过平滑处理和维度统一的处理后，我们得到了初始特征序列。其离散形式函数为。对初始离散序列采用线性拟合技术，我们得到了序列数据的最佳线性拟合连续函数。



**5.2 特征工程**

**5.2.1 衍射峰特征向量**

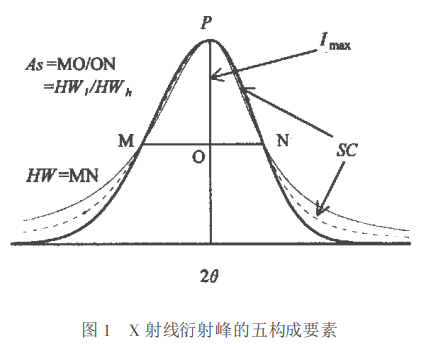
衍射峰，是X射线衍射（XRD）技术中非常重要的概念，其能够反映物质的一些特征，帮助我们进行鉴别和分类。在传统的物相分析中，有许多常用的寻峰算法：对称零面积卷积法、简单比较寻峰法、二阶导数寻峰法、协方差法寻峰、线性拟合寻峰法。本文采用简单高效的线性拟合寻峰法进行寻峰。



**衍射峰个数（）**

对于样本，将其每个峰都记为，那么其第一个特征-峰个数可以被计算：

在峰处，参考文献[X 射线衍射峰五基本要素的物理学意义与应用]，我们针对其几何拓扑特性进行分析并提取了若干个特征。



**衍射峰位置（）**

由线性寻峰拟合法，我们可以确定每个峰的位置，每个峰的寻峰公式为：

其中为线性拟合方法下最逼近原样本曲线的一个多项式函数，我们在后面的实验确定其最佳次数为6次。

**最大衍射峰强度（）**

**衍射峰半高宽（）**

由几何关系，衍射峰的半高宽为：

**衍射峰形态（）**

我们定义衍射峰的形态参数对于对称衍射峰是一个描述除了衍射峰位置、衍射峰高度和衍射峰宽度变化以外的其它变化。这些变化主要表现在衍射峰顶底部的宽缓尖锐程度上的差异, 而造成这种差异的原因就是反映衍射峰形态的物理学意义。

由几何关系，我们定义该参数的计算公式：

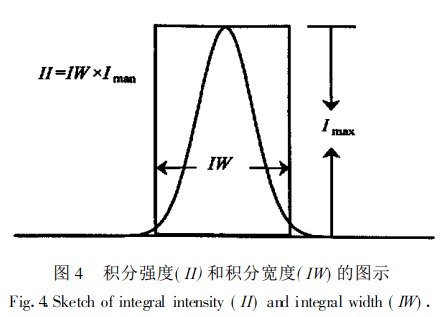
**衍射峰的对称性（）**

**衍射峰积分总强度（）**

其中，[,]是衍射峰所在的区域，边界由一阶导函数的阈值进行确定。

**衍射峰积分宽度（）**

由几何关系，我们定义衍射峰的积分宽度：



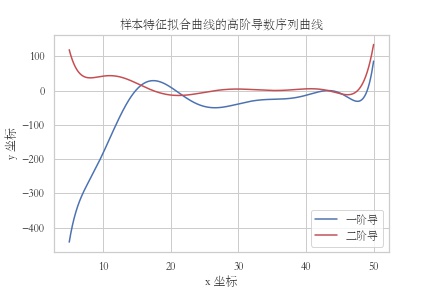
综上，对峰的各种特征做整合，我们得到了曲线的峰特征向量。

**5.2.1 图谱数据高阶导数序列特征**

**高阶导数指纹图谱序列特征**

在对数据进行一定观察后，我们发现数据普遍峰数较少，出现了“模糊性弥散峰”[基于 MATLAB 藏药 X 射线衍射图谱的数据处理]的情况。这种情况会导致峰的一些信息被隐藏，会一定程度影响我们后面的分类和识别。

为了进一步精确地提取这部分的特征，我们对前面线性拟合得到的曲线进行处理，得到指纹图谱数据的一、二阶导数特征序列,。



**5.2.2 图谱数据的特征整合**

综上，对于样本，我们从序列、峰特征两个角度出发抽取特征，对他们做整合，记总的特征向量为：

**5.3 问题一的模型建立与求解**

**5.2.2模型建立**

针对第一问的有监督聚类问题，我们递进地提出了三种方案。

**5.2.2.1 基于单特征序列分布相对熵的分类判别模型**

**样本序列特征**

该模型中，我们只采取了样本数据在滤波后的k=0~2阶的序列导数作为特征向量。

**序列概率归一化处理**

利用样本的序列特征，对相应拟合的曲线函数进行归一化处理后构造了一个概率分布函数 用来唯一表征样本的强度分布。

其中，是一个非常小的常数，用以保证计算KL散度时的分母不为0。

**样本KL散度**

利用概率统计中用来衡量两个分布差异性的KL散度（Kullback-Leibler Divergence），我们将待分类的样本和已分类样本两两进行比较，得到分布之间的KL散度。

**样本最相似类别**

根据待测样本和已标注样本的分布差异数值，我们根据以下公式得到待分类样本的最相似类别。

其中为第c类中的第j个样本的衍射强度概率分布，为第c类所包含的样本个数。

**5.2.2.2 基于多特征相似度投票的分类判别模型**

模型1只采用了单个序列特征，难以充分利用挖掘数据特征。在此模型中，我们采用了更多的特征：峰的几何特征、峰各阶导数序列，能够更充分地发掘图谱数据中的规律。对特征进行分类判别上，我们采用了基于个数值的相似识别算法。

我们的基本思想是：在样本的每个特征点上计算其与已知样本对应特征点上的相似度，并根据相似度确定样本在这个特征点上的最佳分类；对所有特征点进行计算后，我们特征点最佳分类中分类序号出现次数最多，确定样本的综合最佳分类。

**峰、序列特征公式**

**相似度的计算**

在峰个数，峰各阶导数序列这两个维度相等的特征上，我们直接采用了以下的相似度计算方式

其中，为待分类样本和c类第j个样本的距离度量。在后文实验中，我们对三种距离公式：欧式距离、余弦相似度距离、相关系数距离进行测试。

在衍射峰位置（）等这些由于峰个数不定导致维度不定的特征上，无法直接利用距离公式计算，因为可能存在特征维数不等的情况。为了解决这个问题，我们提出了两种解决方法：

1. 我们对样本数据进行了“补峰”操作，在拓展的维度上使用0填充。
2. 取在位置上距离小于一定阈值的峰，其他的峰舍去，以此统一特征维度。

# 图说明

在统一了维度后，这些原本维度不定的特征可以直接采用公式(3-2)的计算方法进行计算。

**每个特征点的最佳分类**

对于样本，在各个特征上，我们通过计算待分类和已知分类样本的相似度，确定在该特征点上样本的最佳分类。

**最佳分类确定**

最后我们根据在特征点上占比最大的类，确定样本的类归属。

**5.2.2.3 基于多阶段特征匹配的概率判别模型**

模型二为基于特征点相似个数的方法，无法考虑各特征上的相似度差距程度和特征重要因子。模型三，通过构造概率下降模型，我们能够将这些因素都充分考虑，实现更加精确的分类。

**分类概率矩阵**

对于样本，我们记其属于类的概率为。由此我们可以构造概率矩阵，在初始时矩阵的每一项=1。

**特征点概率损失**

根据特征重要性，我们构造了一个特征权重向量。这部分作为可学习参数，可以同时扩大模型搜索空间。

采用和模型2统一的相似度计算方案，我们得到样本和类在不同特征点上的相似度矩阵。为了统一不同特征的量纲，我们采用了类似公式(3-2)的归一化方法对各个特征维度进行了处理得到。

记样本与类在特征点上的相似度误差为，我们定义此时的概率损失为。

**概率更新与分类判别**

利用该损失对概率进行更新

最终样本的分类为概率矩阵中对应概率最大的一项确定

**5.2.2 模型求解**

**5.3 问题二的模型建立与求解**

**5.3.1 数据特征向量**

峰特征向量包含峰个数参数，峰几何特征向量。对于不同样本，峰的个数是不定的，所以的维度也是不定的。对于这种特征向量不定的情况，我们无法将样本数据特征向量拼接并统一到等维度的欧式空间去做聚类。

针对这个问题，我们的解决方案是：截取样本区间排序后前k个峰，并对这些峰的特征进行拼接，得到一个统一维度的特征向量。对于峰个数小于k的样本，我们采用了填充的方式。

至此，我们得到了样本的唯一特征向量，其几何意义为n维空间中的一个点。

**5.3.2 模型建立**

针对第二问的无监督学习问题，我们建立了基于Chinese Whisper的无监督聚类模型。

**5.3.2.2 基于Chinese Whisper的聚类模型**

Chinese Whisper是一种基于图的聚类算法，只需要设置一个相似度阀值即可完成无监督分类。

采用第一问模型2的相似度计算方案，我们可以得到相似度矩阵，对其进行遍历构造连接图。

其中，根据第一问的已分类数据拟合得到。

**5.3.2 模型求解**

## 六、模型检验

1. GCN的阈值参数

2. 峰维度统一方式对比

3. 采用基线扣除与否

4.

4.

## 七、模型评价

问题1：

模型1：模型优缺点：简单、直观；仅使用了一个特征，考虑不够全面；计算复杂度较高

模型2：模型优缺点：考虑特征数较多，全面；各特征点的分类重要性被看做一致，缺乏权重分配；算法复杂度更高了

模型3：充分考虑了各特征的差距程度和权重因子，更精确的分类；

问题2：

CW模型优缺点：需要手动调节，相似函数需要人为选定

## 八、模型推广

## 九、参考文献

## 十、附录