

队伍编号	TFB0000502SXJM
题号	A

---

## 基于 MFCC 样本熵和小波分析的地震分类与震级预测研究

### 摘要

随着城市的发展，城建项目急剧增多，地震台监测中出现越来越多的非天然地震事件，对近震事件的记录、大震应急处置与地震目录的日常管理带来了困扰。我们期望在有非天然地震事件干扰等各种客观条件约束下，为地震事件记录处置提供更加便捷有效的处理方案。

本文建立了五个主要模型：模型 I：基于 MFCC 样本熵和免疫遗传算法优化的支持向量机分类模型；模型 II：基于小波分析的最小二乘法线性回归模型；模型 III：水库属性-震级线性回归模型。

针对问题 1，首先对各监测站各地震波形进行分类整合与可视化处理。接着借助 MFCC 对地震波形进行分解提取不同频率特征，再由将 MFCC 静态系数计算出一阶和二阶差分系数从而得出地震波形的动态特征；之后引入样本熵对各阶 MFCC 系数进行特征提取得到特征矩阵进行样本均衡；最后通过免疫遗传算法优化的支持向量机分类模型进行训练，实现地震和爆破分类达到 100% 准确率和精确率。结果见图 11。

针对问题 2，为了根据地震波形的振幅大小和波形特征预测震级，我们先对各个样本进行可视化分析，根据波形振幅的相对大小，剔除了不合理的事件 1、2 和事件 5，并筛选出 13-20 号监测站作为分析样本。之后我们使用 Morlet 小波变换进行特征工程，提取出振幅极差、近似系数能量、细节系数能量等特征；最后通过最小二乘法线性回归模型进行拟合预测，得出附件 9 中的震级为 6.8 级。

针对问题 3，首先对各个特征进行标签编码、归一化等预处理操作，接着我们计算了 Spearman 相关系数并绘制了散点矩阵图来了解特征之间的关系，得出构造活动和震级高度相关；最后我们采用线性回归模型进行回归拟合，得出拟合优度  $R^2=0.976$ ，得出了水库基本属性与震级的关系模型。

最后，对模型进行鲁棒性分析。本文的预测模型在具有轻微扰动的原始输入上的预测结果不会显著偏离该原始输入，说明该模型在输入样本存在细微对抗扰动的情况下，模型预测具有不受对抗样本干扰误导的能力。

关键词：梅尔频率倒谱系数；样本熵；小波分析；震级预测；支持向量机；

# 目 录

<b>第 1 章 问题重述 .....</b>	<b>1</b>
<b>1.1 问题背景 .....</b>	<b>1</b>
<b>1.2 问题描述 .....</b>	<b>1</b>
<b>第 2 章 思路分析 .....</b>	<b>2</b>
<b>2.1 问题 1 的思路分析 .....</b>	<b>2</b>
<b>2.2 问题 2 的思路分析 .....</b>	<b>2</b>
<b>2.3 问题 3 的思路分析 .....</b>	<b>2</b>
<b>第 3 章 符号说明与基本假设 .....</b>	<b>3</b>
<b>3.1 符号说明 .....</b>	<b>3</b>
<b>3.2 基本假设 .....</b>	<b>3</b>
<b>第 4 章 模型的准备 .....</b>	<b>4</b>
<b>4.1 数据预处理 .....</b>	<b>4</b>
<b>4.1.1 数据处理 .....</b>	<b>4</b>
<b>4.1.2 数据可视化 .....</b>	<b>4</b>
<b>4.2 我们的工作 .....</b>	<b>4</b>
<b>第 5 章 问题 1 的模型建立与求解 .....</b>	<b>5</b>
<b>5.1 基于免疫遗传算法的支持向量机分类模型 .....</b>	<b>5</b>
<b>5.1.1 模型的准备 .....</b>	<b>5</b>
<b>5.2 基于免疫遗传算法的求解过程 .....</b>	<b>10</b>
<b>5.2.1 免疫遗传算法求解流程 .....</b>	<b>10</b>
<b>5.2.2 免疫遗传算法流程图 .....</b>	<b>12</b>
<b>5.3 问题 1 的分类结果及分析 .....</b>	<b>12</b>
<b>第 6 章 问题 2 的模型建立与求解 .....</b>	<b>13</b>
<b>6.1 基于小波分析的震级预测线性回归模型 .....</b>	<b>13</b>
<b>6.1.1 模型的准备 .....</b>	<b>13</b>
<b>6.1.2 模型的建立 .....</b>	<b>16</b>
<b>6.2 问题 2 的求解结果 .....</b>	<b>16</b>

<b>第 7 章 问题 3 的模型建立与求解 .....</b>	<b>17</b>
<b>7.1 水库-震级线性回归模型的建立与求解 .....</b>	<b>17</b>
<b>7.2 问题 3 的求解结果 .....</b>	<b>18</b>
<b>第 8 章 模型的测试 .....</b>	<b>18</b>
<b>9.1 鲁棒性分析 .....</b>	<b>18</b>
<b>第 9 章 模型的评价与总结 .....</b>	<b>19</b>
<b>9.1 模型的优点 .....</b>	<b>19</b>
<b>9.2 可能的改进 .....</b>	<b>19</b>
<b>第 10 章 参考文献 .....</b>	<b>19</b>

# 第 1 章 问题重述

## 1.1 问题背景

地震是一种复杂的地壳运动现象，每年全球发生大量的地震灾害事故。为了减少地震灾害，需要开发有效的地震预警预报技术，该技术能够在日常地震监测中准确地识别天然地震事件，并排除人为地震记录或异常干扰信号，然后进行后续处理。精确辨识地震信号是地震学研究和地震观测技术的重要内容。然而，随着城市工程建设项目的增加和地震台网监测范围的扩大，非天然地震事件如爆破、矿震、武器试验和塌陷等越来越频繁发生。这些事件干扰了近震事件的记录、大震应急处置以及地震目录的日常管理。因此，有必要提高地震信号识别模型的可靠性和精度。

针对此问题，我们需要区分出天然地震事件和非天然地震事件。需要挑选合理的事例和样本构建震级预测模型，并准确预测地震等级。另外，我们需要找到水库基本情况和震级的关系，并构建相应的关系模型。

## 1.2 问题描述

**问题 1：**根据附件 1~8 中的地震波数据，找出一系列合适的指标与判据，构建震源属性识别模型，进行天然地震事件（附件 1~7）与非天然地震事件（附件 8）的准确区分。

**问题 2：**地震波的振幅大小、波形特性与震级有着显著关联。根据已知震级大小的附件 1~7 中数据（震级大小分别为：4.2、5.0、6.0、6.4、7.0、7.4、8.0），恰当地挑选事例与样本，建立震级预测模型，并给出附件 9 中地震事件的准确震级（精确到小数点后一位）。

**问题 3：**库深、库容、断层类型、构造活动/基本烈度、岩性等是影响水库诱发地震震级大小的重要因素。根据附件 10 中 102 个水库地震样本，建立水库基本属性资料与震级的关系模型。

## 第 2 章 思路分析

### 2.1 问题 1 的思路分析

问题一要求我们找出合适的指标，构建震源识别分类模型，对天然地震事件和非天然地震事件进行准确区分。根据参考文献和我们自行查阅的文献中得知通常采用支持向量机进行分类，而近年来使用 MFCC 样本熵对地震事件进行分类的研究方法已成为一大热点。首先我们对原始数据进行预处理，对数据进行标准化。在处理过程中并没有发现 NaN 值，最终我们得到长度为 8000 的地震信号数据，并进行可视化处理，整理出每个观测站地震数据的具体情况。然后我们采用 MCFF 梅尔频率倒谱系数和样本熵进行特征工程，从而得到特征矩阵。由于样本数量偏差比较大，我们使用了 SMOTE ENN 组合样本采样方法进行样本均衡。接着建立基于免疫遗传算法优化的支持向量机分类模型，使用免疫遗传算法进行超参数自动寻优，在此基础上使用 MFCC 静态系数样本熵、MFCC 一阶和二阶差分系数样本熵进行训练回归，最终得出预测结果，实现准确区分天然地震事件和非天然地震事件，并对模型的鲁棒性进行分析。

### 2.2 问题 2 的思路分析

问题二要求我们根据地震波的振幅大小、波形特性，从已知震级大小数据中挑选事件和样本，建立模型来预测附件 9 中地震事件的震级。针对此问题，通常根据小波分析得出小波系数、能量和振幅来分析震级之间的关系。我们对已知震级数据预处理后进行 Morlet 小波变换，并可视化各个观测站点的尺度能量图和地震信号折线图，对监测站台和地震事件进行筛选。根据结果进行特征工程，得到包含振幅极差、近似系数能量、细节系数能量和震级的特征矩阵。我们使用线性回归对特征矩阵进行回归拟合，得到震级预测模型，最后对附件 9 中的地震事件进行准确预测。

### 2.3 问题 3 的思路分析

问题三要求我们建立水库与震级的关系模型。针对此问题，我们先对附件 10 进行标签编码、归一化等预处理操作，进行 Spearman 相关系数分析并绘制散点矩阵图进行可视化观察。我们采用线性回归模型进行回归拟合，构建了水库基本属性资料与震级大小的关系模型。

## 第 3 章 符号说明与基本假设

### 3.1 符号说明

表 1：文章中使用的符号及说明

符号	具体说明
$x_n$	第 $n$ 帧的帧信号
$w_n$	窗函数
$s_n$	原始音频信号
$X(k)$	第 $k$ 帧的频谱表示
$H_m(k)$	第 $m$ 个梅尔滤波器在频点 $k$ 的输出
$f[k]$	第 $k$ 个频点的频率
$f_m$	第 $m$ 个滤波器的中心频率
$E_m$	第 $m$ 个滤波器的能量值
$M$	滤波器的数量
$d_t$	第 $t$ 个一阶差分系数
$C_t$	第 $t$ 个倒谱系数
$Q$	倒谱系数 $C_t$ 的阶数
$K$	一阶导数时的差值
$C_n$	第 $n$ 个倒谱系数
$y$	震级
$\beta_i$	第 $i$ 常数项和变量系数
$a_i$	振幅极差、近似系数能量、细节系数能量

### 3.2 基本假设

为了简化问题，我们做出以下基本假设，并且每一个假设都是合理的。

- ✓ **假设 1：**至少有一个监测站点处于震中附近。

**理由：**如果所有的检测站点都处于地震边缘地带，那么加速度、速度和振幅都会与对应震级的特征产生差异，从而导致无法准确识别特征，从而产生预测误差和错误。

- ✓ **假设 2：**各个地震事件之间相互独立。

**理由：**如果各个地震事件存在连锁反应或者连续地震的情况，那地震信号中会存在

时序相关性，对模型准确性造成影响。假设地震事件相互独立，可以使得我们能够使用更多的独立观测数据进行建模和预测。这样的假设可以更好地利用现有的地震数据，并且可以使用不同地震事件的样本进行模型评估和验证。

✓ 假设 3：假设本文研究中使用的数据是准确的。

**理由：**我们假设附件中的数据经过预处理后，与实际情况没有出现明显的测量偏差而被认为是虚假的，因此可以在此基础上建立更合理的量化模型。

## 第 4 章 模型的准备

### 4.1 数据预处理

#### 4.1.1 数据处理

对于附件 1~9 给出的数据，为了提高模型的精度，我们对数据进行预处理，对每个监测站点的数据进行 z-score 标准化，并对各个监测站点记录的数据进行分析。

#### 4.1.2 数据可视化

为了更直观的展示附件 1~7 和附件 9 所给出的数据，方便对问题进行分析，将附件 1~9 中八次自然地震的波形图绘制如下图 1 所示：

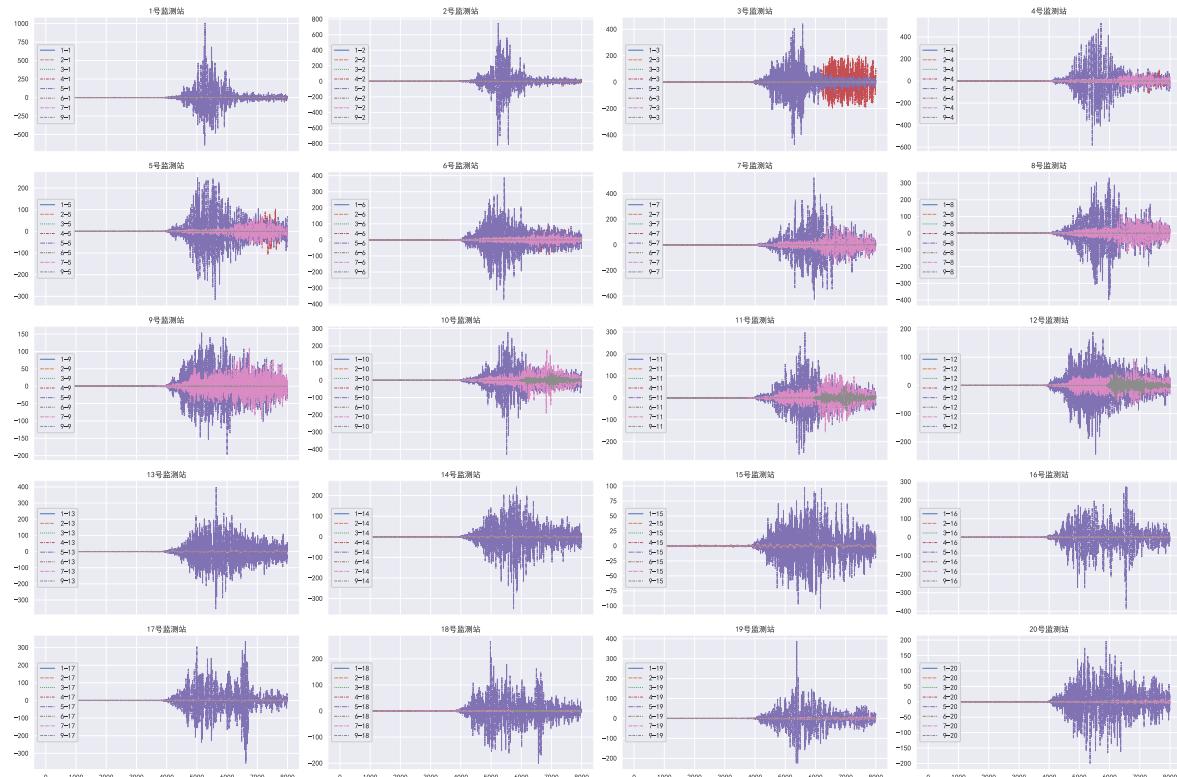


图 1 监测站数据概览图

### 4.2 我们的工作

本文需要我们根据附件 1-7 与附件 8 的数据，建立分类模型对天然地震事件和非天然地震事件进行区分。我们基于 MFCC 样本熵进行了特征工程，并构建了基于免疫遗传算法优化的支持向量机模型进行分类。其次，建立震级预测模型，结合附件 9 数据，根据振幅大小和波形特性预测地震等级。我们使用小波分析对地震波形进行小波变换，得到高低频特征和小波系数作为预测模型的特征，使用线性回归模型进行回归预测。最后，根据附件 10 预处理之后的数据，进行 Spearman 相关系数分析得到震级和水库属性的相关程度，再构建线性回归模型对震级进行预测。综上所述，整个建模过程可以表示成如下图 2 所示：

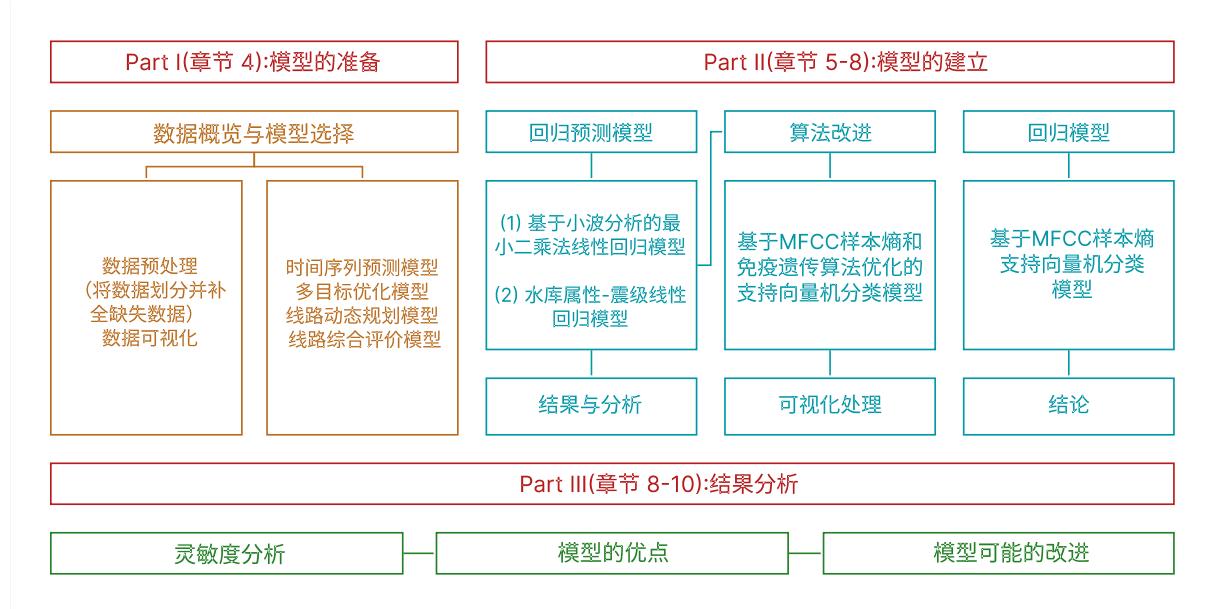


图 2 我们的工作流程图

## 第 5 章 问题 1 的模型建立与求解

### 5.1 基于免疫遗传算法的支持向量机分类模型

#### 5.1.1 模型的准备

##### **Step1:** 计算梅尔频率倒谱系数

MFCC ( Mel Frequency Cepstral Coefficients ) 是一种音频特征提取方法，常用于语音识别和音频处理领域。它是一种基于人耳感知特性的频谱特征表示方法，可以将音频信号转换为一组紧凑且具有判别性的特征向量。MFCC 已经成功应用于语音识别和火山分类。计算过程如下：

(1) 预处理：将原始音频信号分帧，通常每帧持续时间为 20-30 毫秒，并进行加窗处理（例如汉明窗）以减少频谱泄漏。将音频信号分帧其公式如下：

$$x[n] = w[n] \cdot s[n] \quad (1)$$

其中， $x[n]$ 是第 n 帧的帧信号， $w[n]$ 是窗函数， $s[n]$ 是原始音频信号。

(2) 快速傅里叶变换 (FFT)：对每个帧信号应用 FFT 变换，将时域信号转换为频域表示。得到每帧的频谱表示，通常使用复数形式表示频谱，即每个频点包含幅度和相位信息。其公式如下：

$$X(k) = \text{FFT}(x[n]) \quad (2)$$

其中， $X(k)$ 是第  $k$  帧的频谱表示。

(3) 梅尔滤波器组：在频域上定义一组梅尔滤波器，通常为三角形滤波器。梅尔滤波器的中心频率按照梅尔刻度 (Mel scale) 进行设置，以模拟人耳感知频率的非线性特性。对每个频谱点应用梅尔滤波器组，得到每个滤波器的输出能量。其计算方式如下：

$$H_m(k) = \begin{cases} 0 & \text{if } f[k] < f_{m-1} \text{ or } f[k] > f_{m+1} \\ \frac{f[k] - f_{m-1}}{f_m - f_{m-1}} & \text{if } f_{m-1} \leq f[k] \leq f_m \\ \frac{f_{m+1} - f[k]}{f_{m+1} - f_m} & \text{if } f_m \leq f[k] \leq f_{m+1} \end{cases} \quad (3)$$

其中， $H_m(k)$ 是第  $m$  个梅尔滤波器在频点  $k$  的输出， $f[k]$ 是第  $k$  个频点的频率， $f_m$  是第  $m$  个滤波器的中心频率。

(4) 对数压缩：将每个滤波器的能量值进行对数变换，以模拟人耳对音频强度的非线性感知。计算如下：

$$E_m = \log \left( \sum_k |X(k)|^2 \cdot H_m(k) \right) \quad (4)$$

其中， $E_m$ 是第  $m$  个滤波器的能量值。

(5) 倒谱变换：对每个帧的对数能量值序列应用离散余弦变换 (DCT)，DCT 将时域特征转换为频域特征，得到每个帧的倒谱系数。其公式如下：

$$C_n = \sum_{m=1}^M \log(E_m) \cdot \cos \left( \frac{\pi n}{M} (m - 0.5) \right) \quad (5)$$

其中,  $C_n$  是第  $n$  个倒谱系数,  $M$  是滤波器的数量。

(6) MFCC 特征向量提取: 从 DCT 系数中选择前  $N$  个系数作为最终的 MFCC 特征向量。选择前  $N$  个倒谱系数作为 MFCC 特征向量:  $[C_1, C_2, \dots, C_N]$

(7) MFCC 差分系数: MFCC 的计算过程通常只涉及到信号的静态特性。然而, 为了捕捉信号的动态特性, 我们使用一阶差分和二阶差分来提取动态系数。即,

$$d_t = \begin{cases} ccC_{t+1} - C_t, & t < k \\ \frac{\sum_{k=1}^K k (C_{t+k} - C_{t-k})}{\sqrt{2 \sum_{k=1}^K k}}, & \text{其他} \\ C_t - C_{t+1}, & t \geq Q - K \end{cases} \quad (6)$$

该式中:  $d_t$  为第  $t$  个一阶差分系数;  $C_t$  为第  $t$  个倒谱系数;  $Q$  为倒谱系数  $C_t$  的阶数;  $K$  是一阶导数时的差值。

在天然地震中选取附件 5 监测站台 10 的数据为例, 计算得到 MFCC 静态系数、MFCC 一阶差分系数、MFCC 二阶差分系数, 其中信号长度为 8000, 采样率为 200Hz, 离散余弦变换维度为 12, 窗函数采用汉明窗。得到 MFCC 图谱如下图 3、4、5、6 所示。

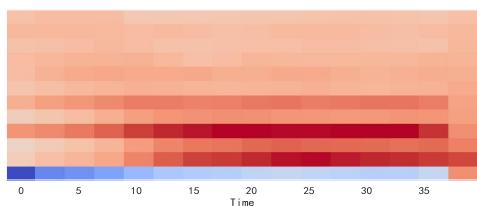


图 3 MFCC 图谱

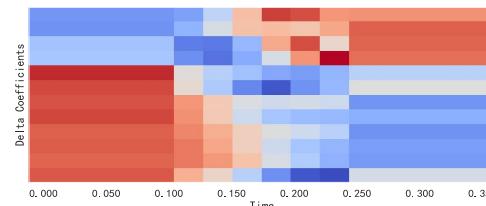


图 4 MFCC 图谱-一阶差分-标准化

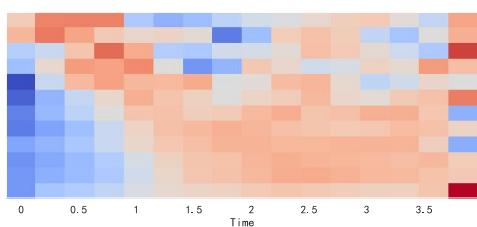


图 5 MFCC 图谱-标准化

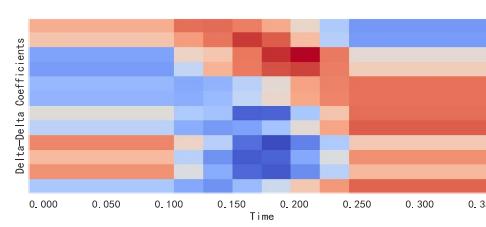


图 6 MFCC 图谱-二阶差分-标准化

对于 MFCC, 低纬度能解释信号整体能量分布和频率特征。而高纬度 MFCC 可以包含更详细和更加丰富的频谱特征, 用于描述信号细微变化。对于地震信号而言, 只需要低维度 MFCC 就能够表达信号的信号能量分部和频率特征。因此提取 MFCC 静态系数、一阶

差分系数和二阶差分系数的第一维度，可视化如图 7 所示。

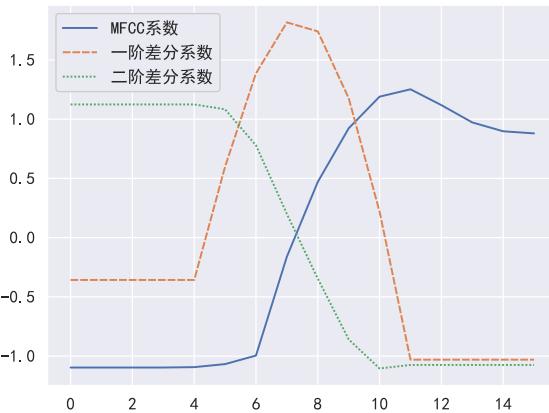


图 7 MFCC 首维系数可视化

### Step2: 计算各阶 MFCC 系数样本熵

样本熵 (Sample entropy, 简称 SampEn) 是 Richman 等在 2000 年提出的一种新的时间序列复杂度表征参数。它是基于近似熵的改进而得到的，这两个指标都用于衡量时间序列的复杂性以及在维度变化时产生新模式的概率大小。如果时间序列产生新模式的概率越大，那么序列的复杂性就越高，其熵值也就越大。样本熵在机械信号分析和故障诊断领域中近年来得到了广泛关注和应用。研究人员们发现，通过计算样本熵，可以有效地分析机械系统产生的信号，并从中提取出有关系统状态和故障的重要信息。这对于预测机械设备的运行状况、实施故障诊断和采取相应的维护措施都具有重要意义。由于非自然地震和自然地震的产生机理、传播方式、波形特征、频率特征等都有明显差异，所以将样本熵应用于 MFCC 各阶系数来描述向量的状态特征。

样本熵的计算步骤为：

(1) 将时间序列  $X$  构造成  $m$  维矢量，即

$$X(i) = \{x(i), x(i+1), \dots, x(i+m-1)\} \quad (7)$$

式中， $i = 1, 2, \dots, N-m+1$

(2) 定义  $X(i)$  与  $X(j)$  间的距离为  $d[X(i), X(j)] (i \neq j)$ ，为两者对应元素中差值最大的一个，即

$$d[X(i), X(j)] = \max_{k \in (0, n=1)} |x(i+k) - x(j+k)| \quad (8)$$

(3) 给定阈值  $r(r > 0)$ , 统计  $d[X(i), X(j)] < r$  的数目并与总的矢量个数  $N - m$  的比值, 即

$$B_i^m(r) = \frac{1}{N - m} \text{num}\{d[X(i), X(j)] < r\} \quad (9)$$

(4) 对所有由式(3)得到的结果求平均, 即

$$B^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} B_i^m(r) \quad (10)$$

(5) 1 再将维数  $m$  加 1, 重复 (1)~(4)

(6) 则理论上此序列的样本熵为:

$$SampEn(m, r) = \lim_{N \rightarrow \infty} \left\{ -\ln \left[ \frac{B^{m+1}(r)}{B^m(r)} \right] \right\} \quad (11)$$

但实际上  $N$  不可能为无穷大, 而为一有限值, 则样本熵的估计值为:

$$SampEn(m, r, N) = -\ln \left[ \frac{B^{m1}(r)}{B^m(v)} \right] \quad (12)$$

根据已经得到的 MFCC 系数, 通过计算得到样本熵特征矩阵, 其中特征分别为 MFCC 静态系数首维向量样本熵、MFCC 一阶差分系数首维向量样本熵、MFCC 二阶差分系数首维向量样本熵, 如图 10 所示。其中,  $x = 140$  左右两侧分别为自然地震和人工爆破。

有图可知, 自然地震和人工爆破特征较为明显, 自然地震由于每次地震中震源位置、地质条件等成分较为复杂, 故样本熵变化明显; 而人工爆破样本熵值变化比较有规律。

### **Step3: 样本均衡**

根据图 8 可知, 人工爆破样本量远小于自然地震样本数量, 这样的样本不均衡会导致分类效果欠佳。因此, 我们使用组合采样 SMOTE ENN 进行样本均衡。从采样前的 170 个样本增加到了采样后的 273 个样本, 其中人工地震样本数量为 138 个, 自然地震样本数量为 135 个。样本均衡后 MFCC 样本熵特征矩阵如图 9 所示。

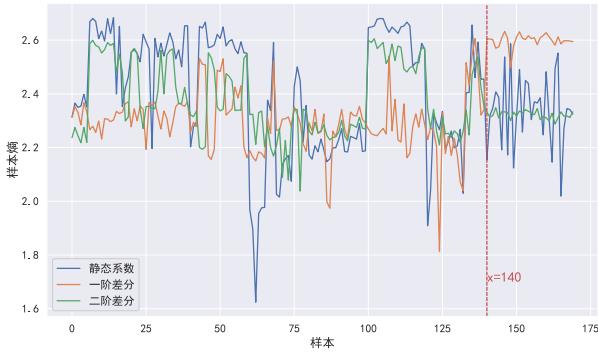


图 8 MFCC 样本熵特征矩阵

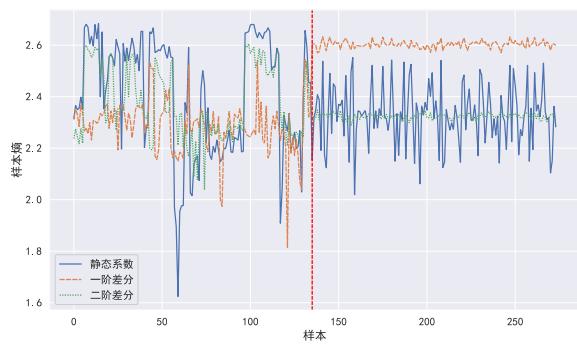


图 9 样本均衡后 MFCC 样本熵特征矩阵

## 5.2 基于免疫遗传算法的求解过程

针对支持向量机等机器学习类模型，超参数的设定对模型结果影响较大，因此采用**启发式算法**进行超参数寻优。传统的遗传算法具有收敛速度较慢，局部搜索能力较弱的缺点，并且运行时间长，容易受参数的影响。而人工免疫算法模拟免疫系统使其具有自我调节能力[8]，以及基于抗体浓度的产生和维持多样性抗体的能力。基于此，我们采用**免疫遗传算法**，将遗传算法和免疫算法的优点结合了起来，通过保存最优个体记忆信息，从而**加速算法的局部搜索能力**，使得进化计算的收敛速度更快，效率更高。

### 5.2.1 免疫遗传算法求解流程

#### **Step1:** 准备

(1) 抗体表示：采用实数编码( $Q_1 \ Q_2 \ \dots \ Q_n$ )，总  $Q_0 = \sum_{k=1}^N Q_k$

其中  $Q_k$  表示第  $k$  个物流场地的货物量 ( $k = 1, 2, \dots, N$ )。 $Q_0$  表示总货物量。

与二进制相比，实数编码在数值优化方面具有更高的精度和效率，并且搜索范围更大。

(2) 适应度  $f$ ：取  $f = Y$

(3) 初始化参数：抗体规模  $N$ ，记忆种群比率  $M$ ，浓度抑制半径  $r$ ，浓度阈值  $T$ ，遗传算子（交叉率  $p_c$ ，变异率  $p_m$ ，灾变率  $p_{m2}$ ）

(4) 终止条件：最大迭代代数  $MAXGEN$

#### **Step2:** 初始抗体群生成

在可行域中，随机产生一个抗体，应用随机模拟，检验满足率的要求，若不满足要求，重新产生一个抗体。直到产生可行的符合满足率的  $N$  个抗体，形成初始抗体群 ( $POP_1, POP_2, \dots, POP_N$ )。

#### **Step3:** 初始抗体群生成

抗体 $v$ 和抗原亲和力 $A_v$ 是指抗体与抗原的匹配度，由目标函数得来。这里采用 $f + 1$ 的倒数。

$$A_v = \frac{1}{1+f} \quad (13)$$

抗体 $u$ 与抗体 $v$ 之间的亲和力 $B_{uv}$ 反映了抗体之间的相似程度，本文采用基于*Euclidean*距离的亲和力计算方法：

$$B_{uv} = \|u - v\| = \sqrt{\sum_{j=1}^N \{u_j - v_j\}^2} \quad (14)$$

两个抗体间的距离在浓度抑制半径 $r$ 内，表示这两个抗体相似程度较高。

#### **Step4: 抗体浓度计算**

抗体浓度 $C_v$ 采用下面公式：

$$C_v = \frac{\sum_{u \neq v} S_{uv}}{N} \quad (15)$$

其中 $N$ 表示抗体个数，

$$S_{uv} = \begin{cases} 1 & \text{若 } \|u - v\| \leq r \\ 0 & \text{其它} \end{cases} \quad (16)$$

在计算抗体浓度时，相似程度比较高的抗体看作是一种抗体。

#### **Step5: 抗体记忆与抑制**

在运算过程中，当一种抗体的浓度超过设定的阈值 $T$ 时，表明抗体在种群中占据了较大优势，达到了一个相对最优点，此时生成一个记忆细胞，以记录此时的局部最优解。每次将亲和度高的部分抗体复制进记忆种群中去，并在每次的迭代中通过记忆种群更新整个种群。

#### **Step6: 遗传**

##### (1) 选择

抗体的选择将抗体群中较好的候选解选择后参与进化。根据适应度公式，得到各个抗体的适应度。按适应度排序，找出抗体群中最大的适应度 $\text{Max } f$ ，定义抗体 $v$ 的浓度， $\alpha$ 、 $\beta$ 为0-1之间的可调参数，可以取一个固定值，可以实现自适应度变化。此模型中取 $\alpha = 0.2$ ， $\beta = 1 - \alpha$ 。采用轮盘赌的方式，选择 $N$ 次，复制 $N$ 个个体到匹配池中等待交叉操作。

##### (2) 交叉

采用顺序算术交叉。根据题目提示DC5关停可能存在不能正常流转的情况，这里交叉算子可能搜索不到靠近边缘的解，应依靠变异。

##### (3) 变异

确定变异个体，随机产生变异方向  $d = (d_1, d_2, \dots, d_N)$ , 其中  $d_i$  是  $Q_i$  的允许变化量

**Step7:** 重复 3-6 步，直到满足终止条件，记忆池中的解为最优解。

### 5.2.2 免疫遗传算法流程图

免疫遗传算法流程图如图 10 所示：

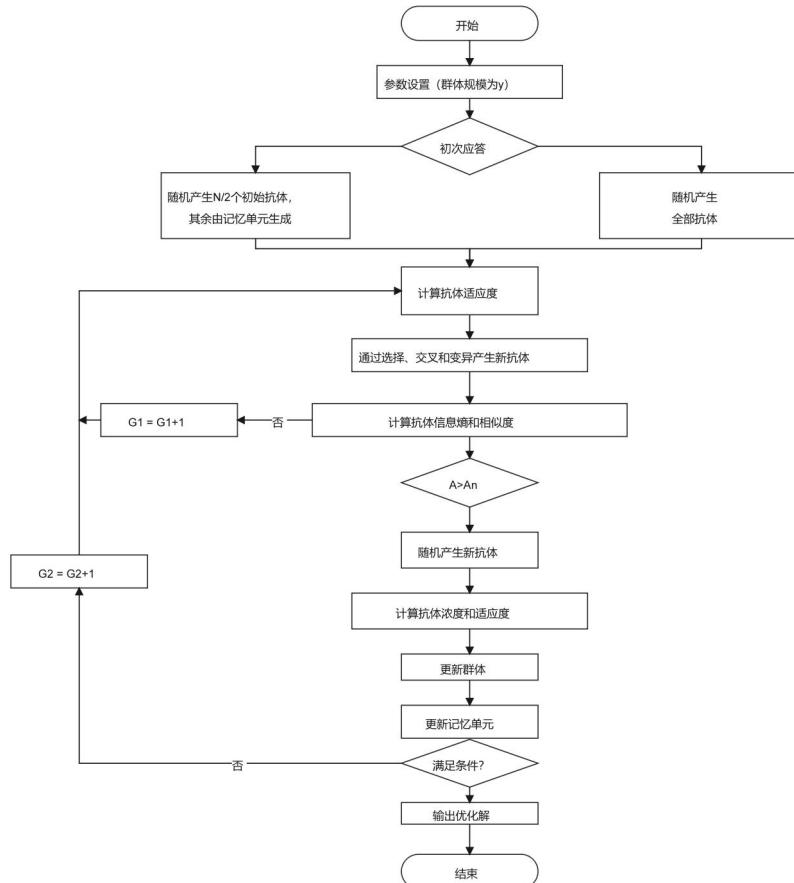


图 10 免疫算法流程图

### 5.3 问题 1 的分类结果及分析

我们将均衡后的 273 个样本使用测试集占比为 0.3 的方法进行划分。通过支持向量机模型得到如图 11 所示的混淆矩阵。

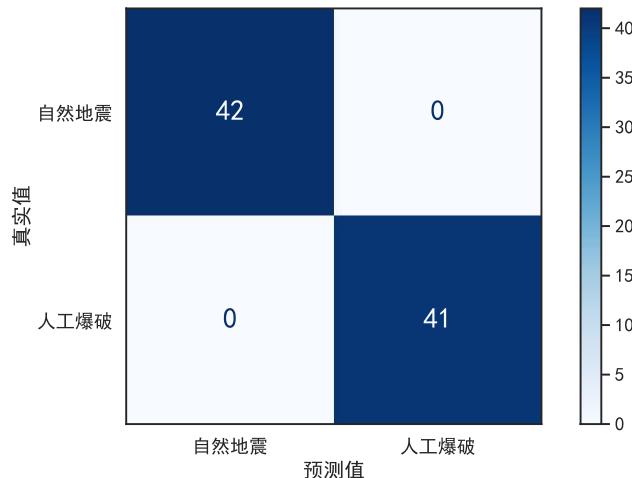


图 11 支持向量机分类结果混淆矩阵

支持向量机模型评估结果如下表 1 所示

表 1 模型评估结果

	准确率	召回率	精确率	F1
训练集	1.00	1.00	1.00	1.00
测试集	1.00	1.00	1.00	1.00

该模型的鲁棒性分析详见第 9 章。

## 第 6 章 问题 2 的模型建立与求解

### 6.1 基于小波分析的震级预测线性回归模型

#### 6.1.1 模型的准备

##### **Step1:** 数据预处理和数据探索

根据问题 2 的相关要求，先对各个地震事件中各个监测站点数据进行可视化分析，挑选出合适的事件和样本。其中各个监测站点历次监测波形图如下图 12 所示。

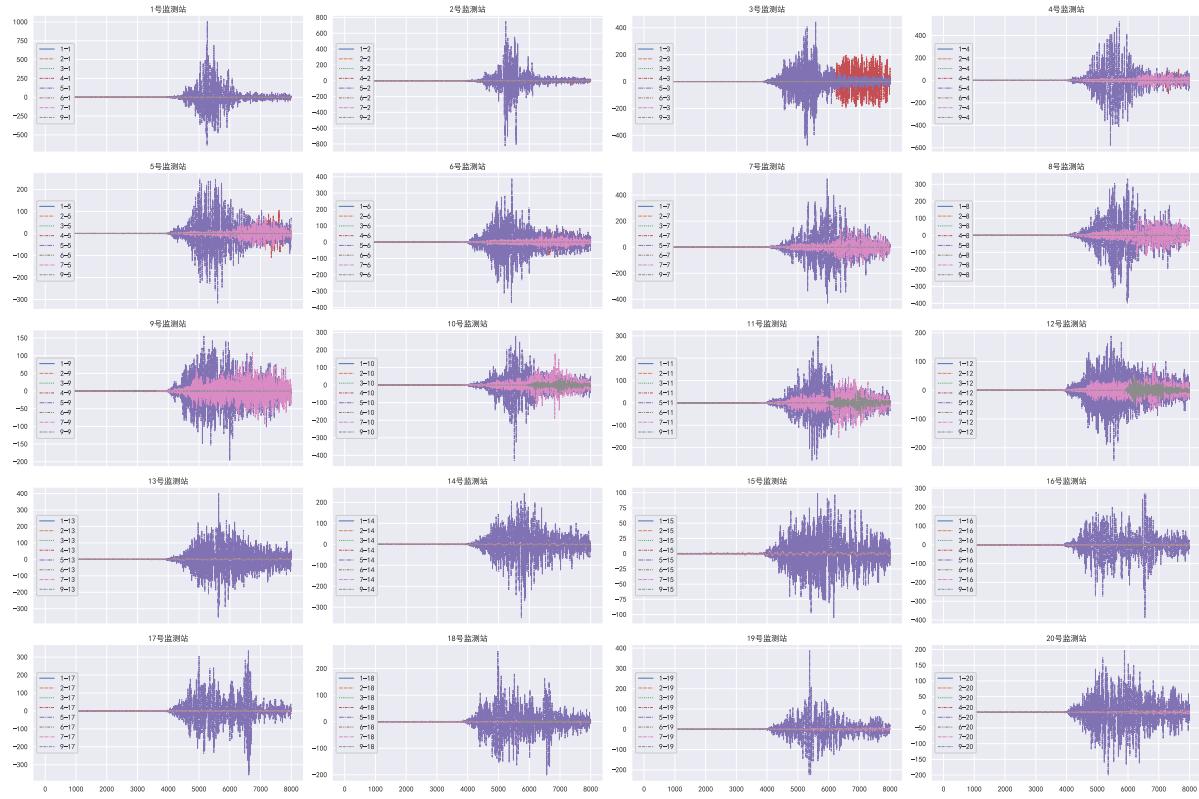


图 12 各个监测站点历次监测波形图

根据上图可视化结果来看，监测数据中存在一个问题：事件 1、2 和 5 的震级小，但是振幅比其他事件都要大，不符合常理，所以应该在样本中剔除事件 1、2 和 5。对事件进行筛选后，得到如图 13 所示波形图。

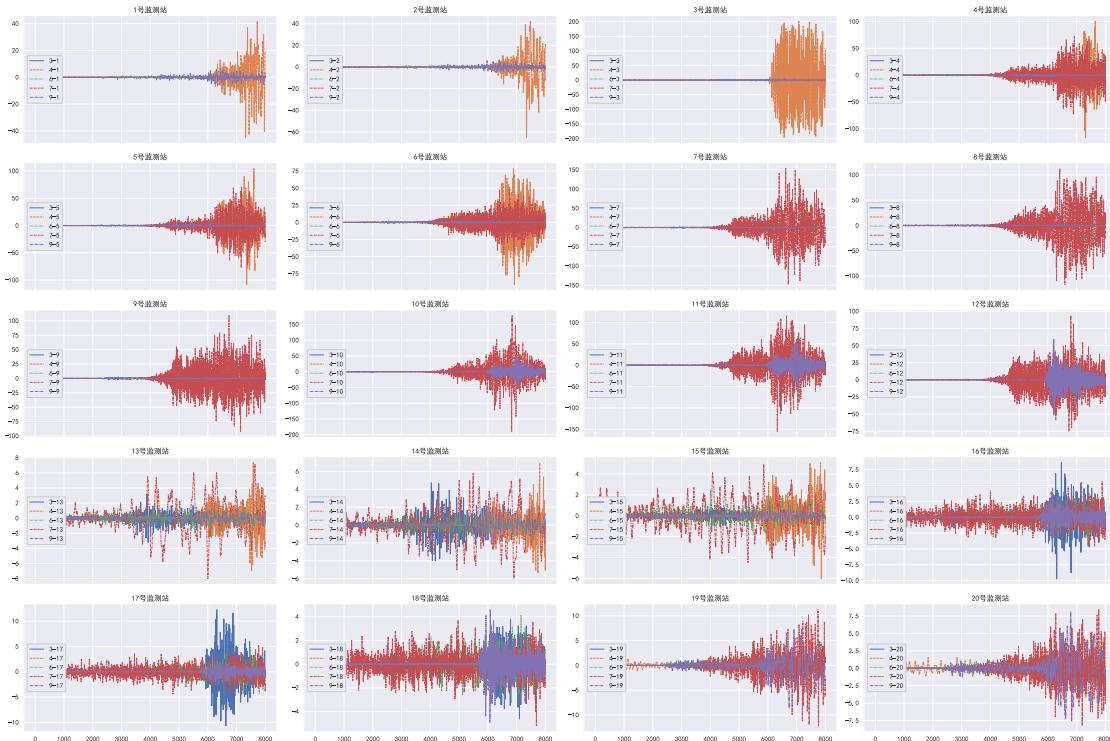


图 13 筛选事件后-各个监测站点历次监测波形图

再次观察波形图，发现 7 号事件所属震级最大，但是在 1-12 号站点都无法明显观察到 7 号事件波形的振幅。因此考虑从样本中剔除 1-12 号站点数据，并保留 13-20 号站点的 3、4、6、7 和 9 号地震事件。

### **Step2: 小波分析和 Morlet 小波变换**

小波分析（Wavelet analysis）是一种数学信号处理方法，用于分析信号的时域和频域特征。它基于小波函数（Wavelet function），通过对信号进行连续或离散的小波变换，将信号分解为不同尺度（频率）的小波成分，从而揭示信号的局部特征和频率信息。

信号  $f(t)$  的连续小波变换  $WT(a, b)$  定义为：

$$WT(a, b) = CWT(a, b) = \langle f(t), \phi_{a,b}(t) \rangle = \iint f(t) |a|^{-\frac{1}{2}} \overline{\phi\left[\frac{t-b}{a}\right]} dt \quad (17)$$

式中，尺度因子  $a$  是指当以离散方式描述某一空间（或时间）函数时，均匀离散点之间的距离； $b$  为位移。

Morlet 小波变换是一种多尺度分析方法，可以在不同时间尺度上对地震信号进行分解。这使得我们可以同时考虑不同频率成分对地震震级的影响，从而获取更全面和详细的信息。公式如下：

$$\phi(t) = \pi^{-1/4} \left( e^{-i\omega_0 t} - e^{-\omega_0^2/2} \right) e^{-t^2/2} \quad (18)$$

对应能量系数  $f(t)$  采用 Morlet 小波进行处理，得到系数矩阵，取所对应的实部的平方再依次求和，公式如下：

$$W_i = \sum_{n=1}^N \operatorname{Re} (E_{in})^2 \quad i = 1, 2, \dots, a \quad (19)$$

式中， $E_m$  为第  $i$  尺度第  $n$  年的小波变换系数。

### **Step3: 特征工程**

第二问要求我们根据波形图预测震级，而波形图无法直接作为输入变量，进行我们需要进行特征工程提取波形特征，对波形图进行合理描述。这一步我们结合小波变换，一共从波形图提取了三个特征，分别为：振幅极差、近似系数能量、细节系数能量。部分数据如下表 2 所示。

表 2: 第 2 问特征工程部分结果

样本	振幅极差	近似系数能量	细节系数能量 0	细节系数能量 1	细节系数能量 2	震级
3-13	0.231995	0.056599	0.031286	0.011377	0.000357	6.0
3-14	0.340815	0.166209	0.065297	0.018936	0.000629	6.0
3-15	0.049270	0.003936	0.010344	0.006490	0.006545	6.0
3-16	0.784573	0.307119	0.724308	0.408168	1.000000	6.0
3-17	1.000000	0.705094	1.000000	0.514291	0.565489	6.0
3-18	0.326524	0.057622	0.559708	0.426304	0.429266	6.0
4-13	0.601240	0.141538	0.573380	0.658489	0.289851	6.4
4-14	0.508403	0.088976	0.442800	0.622714	0.283826	6.4
4-15	0.452280	0.117585	0.381132	1.000000	0.606960	6.4
4-16	0.074209	0.022307	0.000846	0.002633	0.017751	6.4

### 6.1.2 模型的建立

设  $y$  为震级，各个常数项和变量系数为  $\beta_i \in \{1, 2, \dots, 6\}$ ， $a_i \in \{1, 2, 3, 4, 5\}$  依次为振幅极差、近似系数能量、细节系数能量 0、细节系数能量 1、细节系数能量 2

$$y = \beta_1 + \beta_2 * a_1 + \beta_3 * a_2 + \beta_4 * a_3 + \beta_5 * a_4 + \beta_6 * a_5 \quad (20)$$

### 6.2 问题 2 的求解结果

问题 2 模型求解结果如下表 3 所示，线性回归拟合效果图如图 14 所示，拟合优度  $R^2=0.443$ 。事件 9 中监测站点 13-18 的预测结果如下表 4 所示：

表 3: 第 2 问模型求解结果

$\beta$	1	2	3	4	5	6
结果	6. 966	-0. 858	2. 064	-0. 083	0. 238	-1. 263

表 4: 事件 9 中监测站点 13-18 的预测结果

监测站点	13	14	15	16	17	18
预测结果	6. 923602	6. 933214	6. 952266	6. 806296	6. 884583	6. 776356

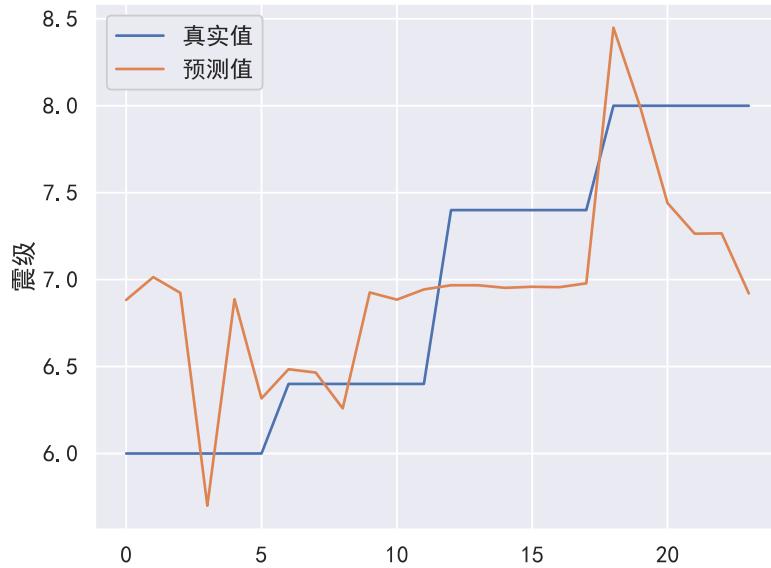


图 14 第 2 问线性回归拟合效果

根据事件 9 中站点的预测结果求平均，得到事件 9 的震级为 6.8 级。

## 第 7 章 问题 3 的模型建立与求解

### 7.1 水库-震级线性回归模型的建立与求解

根据问题 3 的相关要求，先对附件 10 数据进行标签编码。对附件 10 数据进行可视化分析，观察数据分布情况。其中散点矩阵图如下图 15 所示，Spearman 相关系数如图 16 所示。

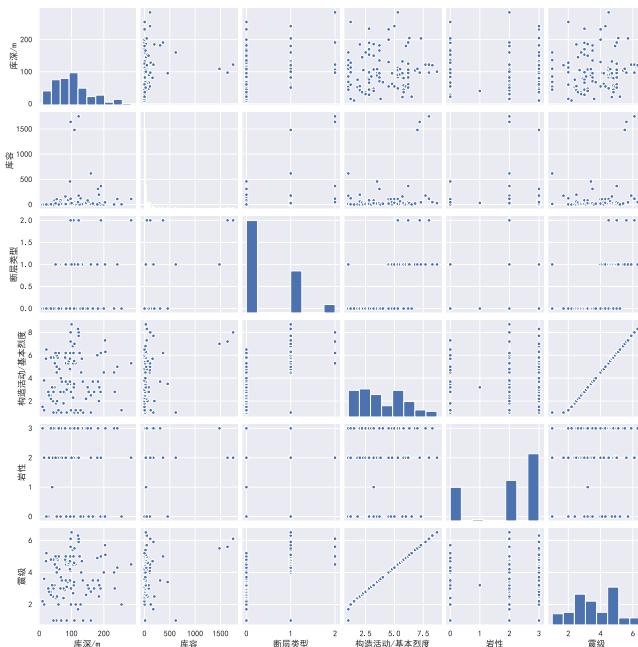


图 15 散点矩阵图

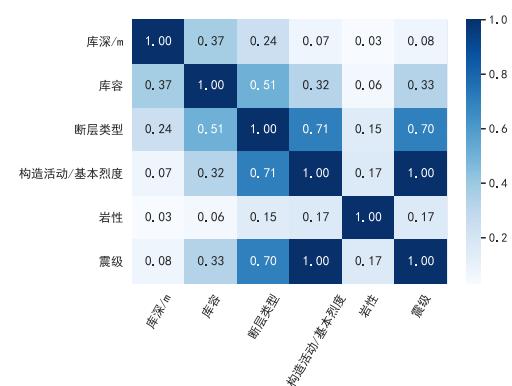


图 16 Spearman 相关系数

## 7.2 问题 3 的求解结果

线性回归拟合效果图如图 17 所示，拟合优度  $R^2=0.976$ 。

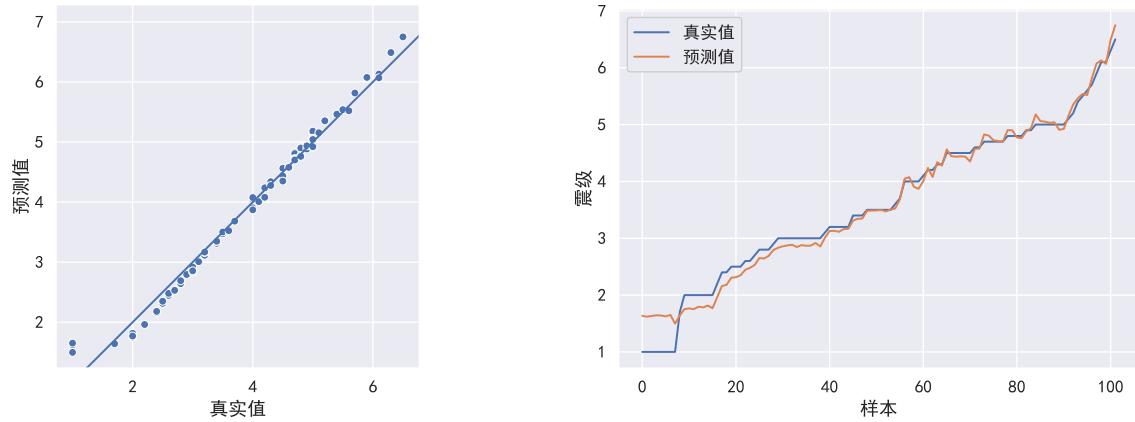


图 17 拟合效果图

## 第 8 章 模型的测试

### 9.1 鲁棒性分析

本文是基于机器学习算法建立模型。机器学习模型的鲁棒性问题会导致模型在环境扰动影响下得到错误的、甚至特定的结果，危害使用者的安全。正确性衡量模型对未知数据的预测性能，即对于给定未知输入，模型能够正常预测或分类的能力；如图 18 所示，本文的预测模型在具有轻微扰动的原始输入上的预测结果不会显著偏离该原始输入，即鲁棒性较强。说明该模型在输入样本存在细微对抗扰动的情况下，模型预测具有不受对抗样本干扰或误导的能力。

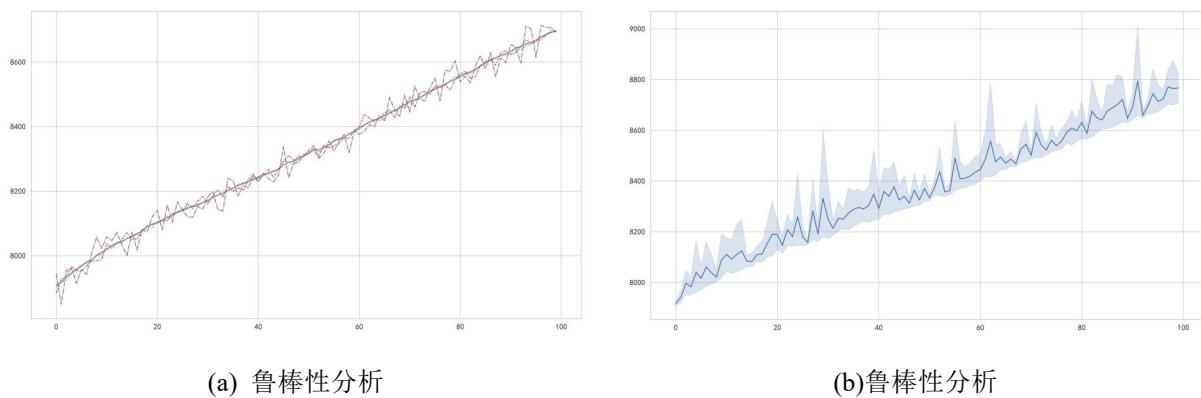


图 18 鲁棒性分析

由此可以看出该模型具有较好的鲁棒性。

## 第9章 模型的评价与总结

### 9.1 模型的优点

- (1) 模型的求解采用启发式算法——免疫遗传算法来自动寻优支持向量机的超参数，将免疫理论和基本遗传算法各自的优点结合在一起，不容易陷入局部最优解；
- (2) 本文基于 MFCC 和样本熵进行分类模型的特征工程相比于直接训练，优化效果显著且准确率极高，对天然地震事件和人工爆破事件辨别与分类提供了有效判据；
- (3) 采用 Spearman 相关系数计算特征相关性，具有较好的客观性；
- (4) 模型的敏感性分析表明了模型在不同参数组合下的有效性，证明了模型的鲁棒性，并且文章可视化效果良好。

### 9.2 可能的改进

- (1) 如果我们有更完整详细的数据，则可以更准确地预测地震震级大小；
- (2) 使用一些近似分析方法对地震震级的预测进行建模，在极端情况下的地震可能会与实际情况相反。

## 第10章 参考文献

- [1] 庞聪,江勇,廖成旺,吴涛,丁炜.基于 MFCC 样本熵和灰狼算法优化支持向量机的天然地震与人工爆破自动识别[J].地震工程学报,2022,44(05):1169-1175.DOI:10.20000/j.1000-0844.20210719002.
- [2] 王晨晖,袁颖,刘立申,陈凯南,吴鹤帅.基于主成分分析法优化广义回归神经网络的地震震级预测[J].科学技术与工程,2022,22(29):12733-12738.
- [3] 吴芳,王卫东,张永志,赵云峰.基于小波分析和最小二乘支持向量机的中国大陆地震震级预测研究[J].地震,2010,30(02):54-60.
- [4] 王博,蒋海昆,宋金.水库诱发地震震级预测的统计研究[J].地震学报,2012,34(05):689-697+727.
- [5] 庞聪,江勇,吴涛,廖成旺,马武刚.神经网络参数对地震类型识别的影响[J].科学技术与工程,2022,22(18):7765-7772.
- [6] 庞聪,丁炜,程诚,吴涛,江勇,马武刚,廖成旺.粒子群优化广义回归神经网络与 HHT 样本熵结合的地震辨识研究[J].地球物理学进展,2022,37(04):1457-1463.
- [7] 范晓易,曲均浩,曲保安等.支持向量分类机 LIBSVM 方法识别天然地震、爆破与塌陷[J].大地测量与地球动力学,2019,39(09):916-918.DOI:10.14075/j.jgg.2019.09.008.
- [8] [1]陈润航,黄汉明,施佳朋等.天然地震与人工爆破地震波形的实时分类研究[J].地球物理学进

展,2019,34(05):1721-1727.

[9] [1]周海军,李磊.地震波形的 HHT 特征提取和 GMM 识别研究[J].黑龙江工业学院学报(综合版),2018,18(04):69-73.DOI:10.16792/j.cnki.1672-6758.2018.04.015.