****    **** 

**中国研究生创新实践系列大赛**

**“华为杯”第二十届中国研究生**

**数学建模竞赛**

|  |  |
| --- | --- |
| **学 校** | **北京石油化工学院** |
| **参赛队号** |  |
| **队员姓名** | **1.邢晓龙** |
| **2.王利猛** |
| **3.张婧** |

**中国研究生创新实践系列大赛**

**“华为杯”第二十届中国研究生**

**数学建模竞赛**

题 目： ­­­­ 基于血性脑卒中临床智能诊疗的建模和优化

摘 要：

注意摘要需包含：建模思路、主要方法、模型、结果与结论、创新点等。摘要内容不超过两页。

出血性脑卒中起病急、进展快，预后较差，病死率高，并且可能会给患者遗留较严重的神经功能障碍，为社会及患者家庭带来沉重的健康和经济负担。研究血肿扩张和血肿周围水肿的发生及发展是针对出血性脑卒中后的两个重要关键事件，所以进行早期识别和预测对于改善患者预后、提升其生活质量具有重要意义。为此我们建立模型，对问题进行分析和研究。

本文所做的工作可概括以下几点：

针对问题一，首先第一小问，

第二小问，采用**随机森林（RF）特征重要度**、**极度梯度提升树（XGBoost）**和**斯皮尔曼相关系数（Spearman）**进行特征筛选，排序投票选出 3个候选集，对候选集取交集，最终确定高血压病史、糖尿病史、冠心病史、止血治疗等变量作为发生血肿扩张的自变量。本文构建**决策树**、**RF**、**Adaboost**、**梯度提升树(GBDT)**等十二类算法，采用 **AUC、Accuracy、Precision、召回率**等多个指标评估， 发现随机森林分类预测模型效果最好，并接着对其参数进行了调优，调优后根据所有含随访影像检查的患者（sub001至 sub100,sub131至 sub160）的90天mRS评分测试集进行预测，平均预测精度 AUC为、Accuracy为、Precision为、召回率为。本文使用随机森林预测血肿扩张预测概率结果见附件“表4”E字段。

第三小问，

问题一：

问题二：

问题三：

针对以上问题设计了各种模型的最优参数组合求解算法，

关键词：

目录

[一、问题重述 3](#_Toc146378606)

[1.1问题背景 3](#_Toc146378607)

[二、问题分析 3](#_Toc146378608)

[2.1 问题一 3](#_Toc146378609)

[2.2 问题二 3](#_Toc146378610)

[2.3 问题三 3](#_Toc146378611)

[三、模型假设 3](#_Toc146378612)

[四、符号说明 4](#_Toc146378613)

[五、问题一模型的建立与求解 4](#_Toc146378614)

[5.1 模型输入的选取 4](#_Toc146378615)

[六、问题二模型的建立与求解 5](#_Toc146378616)

[6.1 模型输入的选取 5](#_Toc146378617)

[6.2 算法流程 5](#_Toc146378618)

[6.3 模型说明 5](#_Toc146378619)

[6.3.1 5](#_Toc146378620)

[6.3.2 5](#_Toc146378621)

[6.4 算法实现 5](#_Toc146378622)

[6.4.1 5](#_Toc146378623)

[6.4.2 5](#_Toc146378624)

[6.4.3 5](#_Toc146378625)

[七、问题三模型的建立与求解 6](#_Toc146378626)

[7.1 算法流程及实现 6](#_Toc146378627)

[7.2 模型说明 6](#_Toc146378628)

[7.2.1 6](#_Toc146378629)

[7.2.2 6](#_Toc146378630)

[7.3.3 6](#_Toc146378631)

[7.3 结果分析 6](#_Toc146378632)

[八、模型的分析与检验 7](#_Toc146378633)

[11.1 误差分析 7](#_Toc146378634)

[11.1.1 问题一的误差分析 7](#_Toc146378635)

[11.1.2 问题二的误差分析 7](#_Toc146378636)

[11.1.3 问题三的误差分析 7](#_Toc146378637)

[11.2 模型的检验 7](#_Toc146378638)

[11.3模型的不足 7](#_Toc146378639)

[九、模型的评价 7](#_Toc146378640)

[12.1模型优点 7](#_Toc146378641)

[12.2模型缺点 7](#_Toc146378642)

[十、参考文献 8](#_Toc146378643)

[附录源程序 8](#_Toc146378644)

# 一、问题重述

## 1.1问题背景

出血性脑卒中指非外伤性脑实质内血管破裂引起的脑出血，占全部脑卒中发病率的10-15%。其病因复杂，通常因脑动脉瘤破裂、脑动脉异常等因素，导致血液从破裂的血管涌入脑组织，从而造成脑部机械性损伤，并引发一系列复杂的生理病理反应。出血性脑卒中起病急、进展快，预后较差，急性期内病死率高达45-50%，约80%的患者会遗留较严重的神经功能障碍，为社会及患者家庭带来沉重的健康和经济负担。因此，发掘出血性脑卒中的发病风险，整合影像学特征、患者临床信息及临床诊疗方案，精准预测患者预后，并据此优化临床决策具有重要的临床意义。

出血性脑卒中后，血肿范围扩大是预后不良的重要危险因素之一。在出血发生后的短时间内，血肿范围可能因脑组织受损、炎症反应等因素逐渐扩大，导致颅内压迅速增加，从而引发神经功能进一步恶化，甚至危及患者生命。因此，监测和控制血肿的扩张是临床关注的重点之一。此外，血肿周围的水肿作为脑出血后继发性损伤的标志，在近年来引起了临床广泛关注。血肿周围的水肿可能导致脑组织受压，进而影响神经元功能，使脑组织进一步受损，进而加重患者神经功能损伤。综上所述，针对出血性脑卒中后的两个重要关键事件，即血肿扩张和血肿周围水肿的发生及发展，进行早期识别和预测对于改善患者预后、提升其生活质量具有重要意义。

医学影像技术的飞速进步，为无创动态监测出血性脑卒中后脑组织损伤和演变提供了有力手段。近年来，迅速发展并广泛应用于医学领域的人工智能技术，为海量影像数据的深度挖掘和智能分析带来了全新机遇。期望能够基于本赛题提供的影像信息，联合患者个人信息、治疗方案和预后等数据，构建智能诊疗模型，明确导致出血性脑卒中预后不良的危险因素，实现精准个性化的疗效评估和预后预测。相信在不久的将来，相关研究成果及科学依据将能够进一步应用于临床实践，为改善出血性脑卒中患者预后作出贡献。

1.2问题提出

基于附件数据与出血性脑卒中知识背景，本文需要解决下列三大问题：

**问题1. 血肿扩张风险相关因素探索建模**

a) 根据附件“表1”和“表2”，判断患者sub001至sub100发病后48小时内是否发生血肿扩张事件。如发生血肿扩张事件，同时记录血肿扩张发生时间到附件“表4”D字段。

b）以是否发生血肿扩张事件为目标变量，基于附件“表1” 前100例患者的个人史，疾病史，发病相关、附件“表2”中其影像检查结果及附件“表3”其影像检查结果等变量，构建模型预测所有患者（sub001至sub160）发生血肿扩张的概率于附件“表4”E字段。

**问题2. 血肿周围水肿的发生及进展建模，并探索治疗干预和水肿进展的关联关系**

a）根据附件“表2”前100个患者的水肿体积和重复检查时间点，构建一条全体患者水肿体积随时间进展曲线，计算前100个患者真实值和所拟合曲线之间存在的残差于附件“表4”F字段。

b）探索患者水肿体积随时间进展模式的个体差异，构建不同人群的水肿体积随时间进展曲线，并计算前100个患者真实值和曲线间的残差于“表4”G字段，所属亚组填写在H段。结

1. 请分析附件“表1”中不同治疗方法对水肿体积进展模式的影响。
2. 请分析附件“表1”中血肿体积、水肿体积及治疗方法三者之间的关系以及出血性脑卒中患者预后预测及关键因素探索。

**问题3. 出血性脑卒中患者预后预测及关键因素探索**

a）根据附件“表1”中前100个患者个人史、疾病史、发病相关及附件“表2”、“表3”中首次影像结果构建预测模型，预测前100个患者90天mRS评分于附件“表4”I字段。

b）根据附件“表1”前100个患者所有已知临床、治疗附件“表2”及“表3”的影像（首次+随访）结果，预测所有含随访影像检查的患者（sub001至sub100,sub131至sub160）90天mRS评分于附件“表4”J字段。

c）分析出血性脑卒中患者的预后（90天mRS）和个人史、疾病史、治疗方法及影像特征（包括血肿/水肿体积、血肿/水肿位置、信号强度特征、形状特征）等关联关系，为临床相关决策提出建议。

# 二、问题分析

## 2.1 问题一

a)

b) 问题b）要求依据“表1” 前100例患者（sub001至sub100）的个人史，疾病史，发病相关（字段E至W）、“表2”中其影像检查结果（字段C至X）及“表3”其影像检查结果（字段C至AG）等变量，构建模型预测所有患者（sub001至sub160）发生血肿扩张的概率。，本文从以下三个步骤解决问题二：（1）对附件“表1”、“表2”、“表3”中对数据缺失值较多删除，并且采用割线法对缺失值较少的进行补全；（2）对附件“表1”、“表2”、“表3” 中特征变量进行筛选，采用随机森林（RF）特征重要度、极度梯度提升树（XGBoost）和斯皮尔曼相关系数（Spearman）先剔除冗余的强相关性变量，再按特征重要性排序，然后采用投票方式选出 3组分别对发生血肿扩张具有显著影响的变量候选集，最后将 3 组候选集取交集后得到最终筛选出的特征变量。（3）十二个模型进行预测对比择优后，选择最优模型-随机森林优化参数干预测发生血肿扩张的概率。

问题一（b）的总体思路如图 5.1 所示

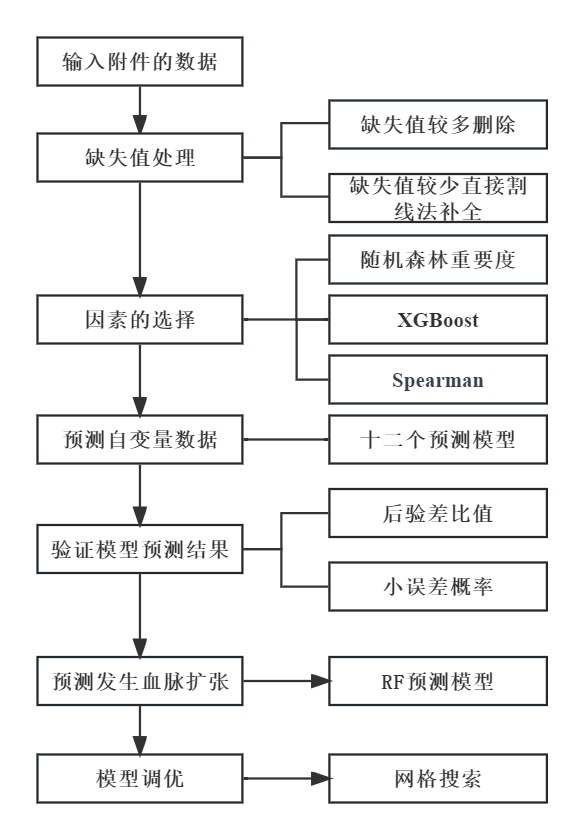


图 问题一b）的解题思路

## 2.2 问题二

## 2.3 问题三

# 三、模型假设

1.问题1假设：

2.问题2假设：

3.问题3假设：

# 四、符号说明

本文所使用的符号系统及其解释如表 4.1 所示。

表 4.1 本文所使用的部分符号说明

|  |  |
| --- | --- |
| 符号 | 符号说明 |
| Y | 参考数列 |
| Xi | 比较数列 |
| Xrescaled | 无量纲化 |
|  | 灰色关联系数 |
| yi | 预测输出 |
| oi | 期望输出 |
| E | 误差绝对值和 |
| F | 个体适应度 |
| pi | 选择概率 |
| i | 个体 |
| Fi | 个体的适应度值 |
| N | 种群个体数目 |
| ak | 第k个染色体 |
|  | 第k个染色体ak在j位的交叉操作 |
|  | 第l个染色体a1在j位的交叉操作 |
| amax | 基因aij的上界 |
| amin | 基因aij的下界 |
|  | 预测值 |
|  | 期望值 |
| T | 最大迭代次数 |
| L | 损失函数 |
| ctj | 最佳负梯度拟合值 |

注：考虑到全文连续性，其他未在表 4.1 中列出的符号将在建模和求解过程中给出解释说明。

# 五、问题一模型的建立与求解

## 5.1 问题a）分析

5.1.1数据缺失值处理

对数据缺失较多的变量进行删除，对缺失值较少的数据采取割线法进行补全，迭代式如下：

保留的变量如表5.1所示。

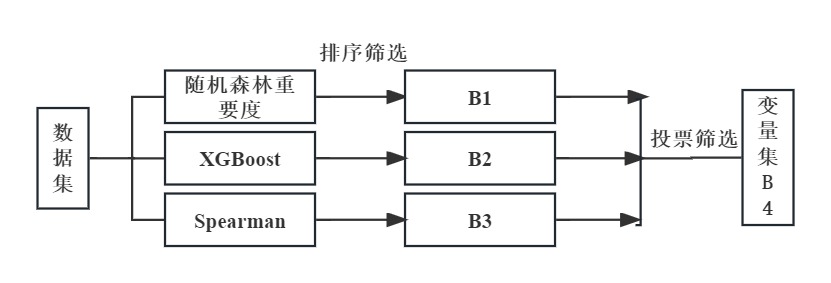
表5.1 保留的变量

|  |  |  |
| --- | --- | --- |
| HM\_volume0 | original\_shape\_Sphericity | 高血压病史 |
| HM\_MCA R\_Ratio | Original\_shape\_SurfacevolumeRatio | 糖尿病史 |
| HM\_PCA R\_Ratio | NccT\_original\_firstorder\_10Percentile | 冠心病史 |
| HM\_PCA\_L\_Ratio | NCCT\_original\_firstorder\_InterquartileRange | 止血治疗 |
| ED volume | NCCT\_original\_firstorder\_MeanAbsoluteDeviation | 降颅压治疗 |
| ED\_MCA\_R\_Ratio | NCCT\_original\_firstorder\_Mean | 降压治疗 |
| ED\_MCA\_L\_Ratio | NCCT\_original\_firstorder\_Median | 年龄 |
| original\_shape\_Elongation | NccT\_original\_firstorder\_Minimum | 性别 |
| original\_shape\_Flatness | NcCT\_original\_firstorder\_Range | 脑出血前mRS评分 |
| original\_shape\_Maximum2DDiameterColumn | NCCT\_original\_firstorder\_RobustmeanAbsoluteDeviation | 发病到首次影像检查时间间隔 |
| original\_shape\_Maximum2DDiameterRow | NCCT\_original\_firstorder\_RootmeanSquared | 收缩压 |
| original\_shape\_Waximum2DDiameterSlice | NCCT\_original\_firstorder\_Skewness | 舒张压 |
| original\_shape\_MinorAxisLength | NCCT\_original\_firstorder\_Variance |  |

5.1.2 特征筛选

通过对变量进行筛选，找出对影响发生血肿扩张最显著的变量。通过3个方法，可以对所有变量进行近似表达。各变量间的作用可以用相关性来表达，且相关性越大，其作用愈明显。由于传统的特征筛选法很难精确地反映各个变量间的相关性，而且各个方法的原理及衡量结果也不尽相同，所以本文采用了三种包括线性和非线性特征筛选法的综合选取，并将三种方法中得到的变量相关性排序结果进行了综合，从中得到最好的特征变量，并给出了如下的特征筛选模型：

假定在方法 1 至 2 中，所选的变量集为 B1、 B2、 B3，各个变量集包含多个变量，且根据变量的相关性程度进行排序。选取最佳变量的指标为（1）变量出现频数（愈多愈好）（2）变量排序（愈往前愈好）。由于各项方法的评估结果都存在差异，所以本文选取了人工的投票模型选择变量。最后将 3 个 B5 集合中所得出的变集取交集得出影响发生血肿扩张的“公有变量”为最终结果。



1. XGBoost

XGBoost 是 boosting 算法的一种实现方式，能够有效减少模型的误差。基本思路为不断生成新的决策树，每棵树都是基于上一颗树和目标值的差值来进行学习，从而降低模型的偏差。

式中，n 是样本数。i表示样本的索引。ℓ是损失函数，用于衡量预测值与实际值 之间的差异，​是正则化项，用于控制模型的复杂度，避免过拟合。是模型的参数，包括每棵树的结构和叶子节点上的预测值。

1. 随机森林重要度

随机森林特征重要度的计算通常基于两个主要指标：Gini重要度和平均减少（Mean Decrease Impurity）。

Gini重要度是通过评估每个特征在决策树节点上的Gini不纯度减少量来计算的。在随机森林中，Gini重要度是对每个特征的Gini减少量的平均值。

计算公式：

式中，T是树党的数量，是树t 上的Gini不纯度减少量，Weight(t)是树t 的权重（通常为叶子节点样本数或节点样本数）。

平均减少不纯度衡量了每个特征在每个决策树上降低的不纯度的平均值。该指标用于确定每个特征对模型准确性的贡献。

计算公式：

式中，Impurity\_parent (t)表示树 t 的父节点的不纯度，Impurity\_left(t)和Impurity\_right(t)分别表示树 t 的左子节点和右子节点的不纯度。

（3）Spearman 相关系数分析

斯皮尔曼相关系数（Spearman Correlation Coefficient）是一种用于度量两个变量之间的单调关系（可能是非线性的）的统计量。它不要求变量是连续的，可以是有序变量或者连续变量。斯皮尔曼相关系数是通过计算变量的秩（rank）来确定相关性的。

公式如下：

式中n 是样本数量， rs=1 表示完全正相关。rs =−1 表示完全负相关。rs =0 表示无相关关系

（4）排序，投票，交集

通过对 3种不同方法筛选出的变量集进行比较，我们可以看出，不同的方法可以得到不同的特征集，但也有部分变量在每个变量集中重复出现。将每个变量集中的变量按照其是否出现分别赋值 1（出现）和 0（未出现），选择那些在多个变量集中值为 1 的变量进入候选集，对发生血肿扩张对应的3个候选集取交集，最终交集包含的筛选出的变量如表 5.4 所示。

5.1.3 对比分析与模型验证

本文对“表1”、“表2”和“表3”的数据共 100 条按 8：2划分训练集和测试集，统一采用 5 折交叉验证的方式对包括 MLP、 Decision Tree、 AdaBoost 等 11 个回归预测模型进行验证，检验其平均精度。采用AUC为、Accuracy为、Precision为、召回率作为 4 个评价指标来确定最优回归预测模型。

以 10cm 湿度(kg/m2)的模型为例（其他深度的土壤湿度预测实验对比效果类似），

各个模型交叉验证的最终实验结果如图 5.11 所示。设置合适的参数对机器学习方法

十分重要，为了保证实验结果的客观性和可对比性，这里用到的模型均采用默认参

数。

5.1.4 参数调优及求解

## 六、问题二模型的建立与求解

问题2由4小问组成。

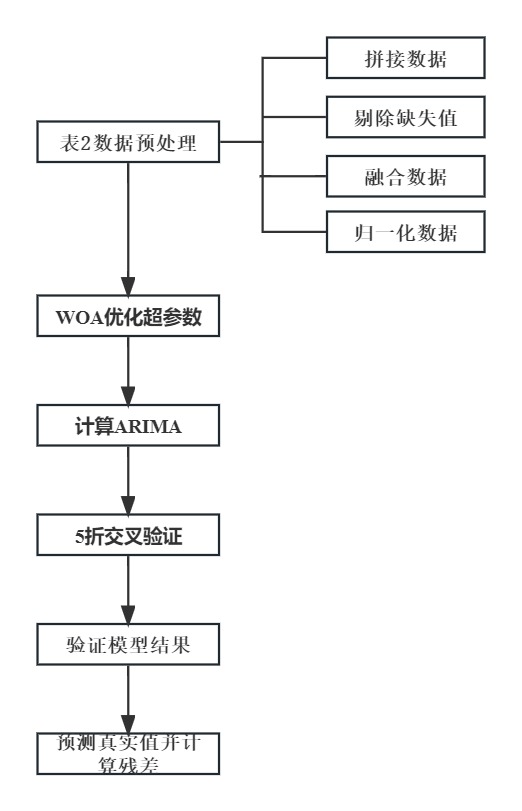
其中，问题a属于模型拟合类型。要求分析全体患者水肿体积与时间之间的关系。难点在于该时间是否应该考虑多次随访记录。假设首次影像检查是在治疗前，第一次随访检查是在治疗后。第一次随访所检测到的水肿体积势必会受到治疗方式等外部因素的影响，无法与时间保持唯一相关。由于每一名患者从发病到首次影像检查时间间隔不尽相同，且相邻随访时间间隔不固定。这就导致即使是在同一个时间点，有的患者可能已经接受过治疗，有的则可能尚未进行首次影像检查。即固定的时间点对不同人是否已经接受过治疗具有歧义性，而是否接受过治疗会直接影响到水肿体积。因为治疗前后水肿体积值不是同一量级，所以无法通过简单的平均去融合固定时间点的水肿体积值。综上，为了避免这种风险，选择研究发病到首次影像检查时间间隔与水肿体积的关系。因为患者从发病到首次影像检查未经过治疗，水肿体积仅由患者个人信息决定，且后者不随时间对水肿体积产生影响，此时水肿体积仅与时间相关。考虑到自变量是时间，可以构建并训练ARIMA时序模型，用WOA算法优化超参数并用5折交叉验证说明模型质量，最后计算残差。

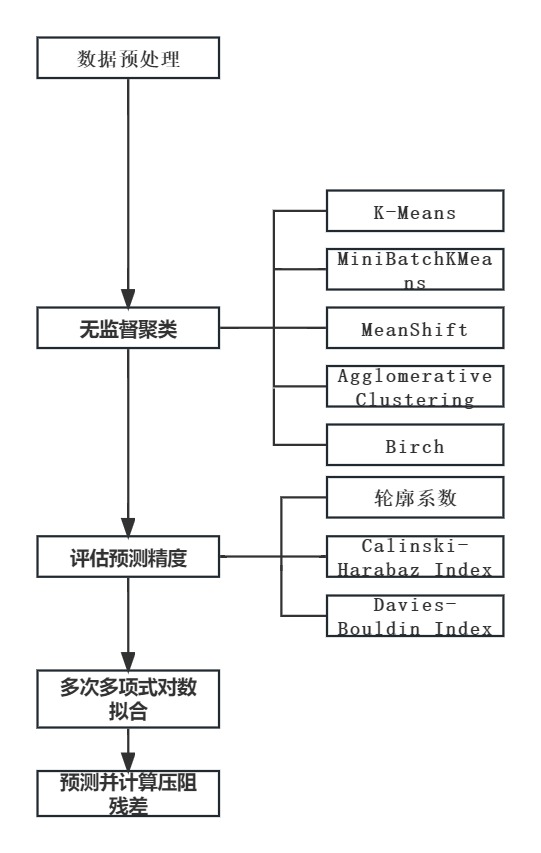
问题b与a类似，因而也选择发病到首次影像检查时间间隔作为时序自变量。不同之处在于，需要先选择特征对患者聚类分组，再训练若干模型分别计算亚组残差。首先根据肘部法确定合适的组数，在聚类算法的选择上，使用了KMeans,MiniBatchKMeans,SpectralClustering,AgglomerativeClustering,Birch共5种聚类算法，同时用轮廓系数，Calinski-Harabaz Index和Davies-Bouldin Index作为评估标准，选择最佳的AgglomerativeClustering聚类结果。对每一类亚组都分别使用1至5次多项式、指数和对数拟合。根据R2系数选择最佳拟合方式并计算亚组残差。

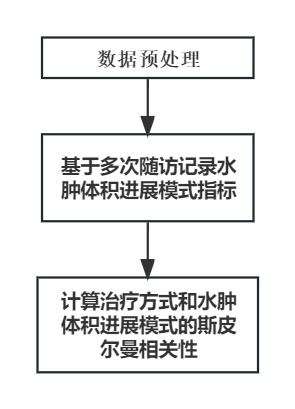
问题c属于相关性分析，难点在于水肿体积进展模式特征的确定。定义相邻随访记录单位时间内水肿体积变化量为此特征，计算治疗方法与其相关性。

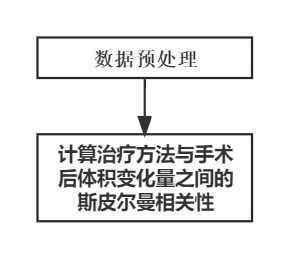
问题d与c相似，需要注意的是应该研究患者接受治疗后，治疗方式与血肿体积和水肿体积的相关性。因而直接计算治疗方式与多次随访的血肿体积和水肿体积的相关性即可，不必考虑首次检查的血肿体积和水肿体积。

## 6.1 算法流程







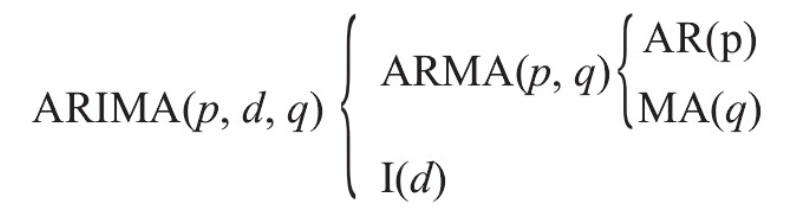


## 6.2 模型说明

### 6.2.a

#### ARIMA：

本问中，使用差分自回归平移(Autoregressive Integrated Moving Average，ARIMA)模型来研究发病到首次影像检查时间间隔与水肿体积的关系。ARIMA可以有效地预测时间序列。如果时间序列具有趋势，则对其作差分后变为平稳随机序列，再用平稳时间序列去描述这一随机过程。运用最佳拟合的模型，对过去、现在的时间序列观测值对未来数据进行预测[1]。ARIMA(p,d,q)结构关系如下：



模型结合了三种基本方法：

* 自回归（AR） - 在自回归的一个给定的时间序列数据在他们自己的滞后值，这是由在模型中的“P”值表示回归的值。
* 差分（I-for Integrated） - 这涉及对时间序列数据进行差分以消除趋势并将非平稳时间序列转换为平稳时间序列。这由模型中的“d”值表示。如果d = 1，则查看两个时间序列条目之间的差分，如果d = 2，则查看在d = 1处获得的差分的差分，等等。
* 移动平均线（MA） - 模型的移动平均性质由“q”值表示，“q”值是误差项的滞后值的数量。

ARIMA模型对数据的处理方式如下：

*第1步：检查数据是否平稳且相关*

使用Augmented Dickey-Fuller单位根测试测试平稳性。对于平稳的时间序列，由ADF测试得到的p值必须小于0.05或5％。如果p值大于0.05或5％，则可以得出结论：时间序列具有单位根，这意味着它是一个非平稳过程。为了将非平稳过程转换为平稳过程，应用差分方法。将差分值形成新的时间序列数据集。可以连续多次应用差分方法，产生“一阶差分”，“[二阶差分](https://www.zhihu.com/search?q=%E4%BA%8C%E9%98%B6%E5%B7%AE%E5%88%86&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7b)”等。应用适当的差分顺序（d）使时间序列平稳。

*第2步：确定p d q*

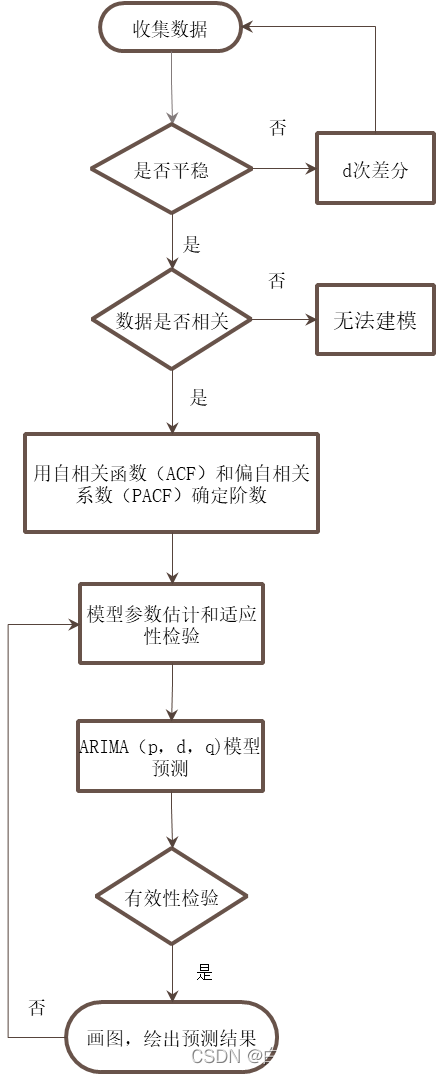
通过[自相关函数](https://www.zhihu.com/search?q=%E8%87%AA%E7%9B%B8%E5%85%B3%E5%87%BD%E6%95%B0&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra={"sourceType":"answer","sourceId":1595121601}" \t "_blank)（ACF）和偏相关函数（PACF）来确定自回归（AR）和移动平均（MA）过程的适当顺序，从而确定p,d,q。对于AR模型，ACF将以指数方式衰减，PACF将用于识别AR模型的顺序（p）。如果在PACF上的滞后1处有一个显着峰值，那么就会有一个[1阶AR模型](https://www.zhihu.com/search?q=1%E9%98%B6AR%E6%A8%A1%E5%9E%8B&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7b)，即AR（1）。如果在PACF上有滞后1,2和3的显着峰值，那么就会有一个[3阶AR](https://www.zhihu.com/search?q=3%E9%98%B6AR&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7b)模型，即AR（3）；对于MA模型，PACF将以指数方式衰减，ACF图将用于识别MA过程的顺序。如果在ACF上的滞后1处有一个显着的峰值，那么就会有一个1阶的MA模型，即MA（1）。如果在ACF上的滞后1,2和3处有显着的峰值，那么就会有一个3阶的MA模型，即MA（3）。

*第3步：估算和预测*

一旦确定了参数（p，d，q），就可以估算ARIMA模型在训练数据集上的准确性，然后使用[拟合模型](https://www.zhihu.com/search?q=%E6%8B%9F%E5%90%88%E6%A8%A1%E5%9E%8B&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra={"sourceType":"answer","sourceId":1595121601}" \t "_blank)使用预测函数预测测试数据集的值。

*第4步：交叉验证*

最后，我们交叉检查我们的预测值是否与实际值一致。



ARIMA流程图 按上文的步骤修改下流程图[3]

#### WOA算法 公式：

WOA（鲸鱼优化算法）旨在解决传统优化算法在处理复杂问题方面的局限性。WOA主要包括：包围猎物，泡网攻击方法，搜索猎物三步。

1.在包围猎物阶段，WOA算法将一个人设置为最佳解决方案，其他个体根据最优解更新其位置。此行为由公式 （9） 和公式 （10） 表示。[3]

式中:t为当前迭代次数，、为系数向量，X \* (t)为当前最优解的位置向量，X(t)向量为位置向量。上式中计算为:

式中:a从2逐渐减小到0,r1和r2为[0,1]中的随机向量。

2.在气泡网攻击方法阶段，设计了两种方法。收缩包围圈机制:这种方法类似于包围圈猎物阶段，公式没有太大区别，只是的取值范围被限制在−1到1之间。螺旋更新位置:通过在鲸鱼(X, Y)和猎物(X\*，Y\*)位置之间创建一个螺旋方程，数学模型如下:

式中:L = |X \* (t)−X(t)|表示鲸鱼与其猎物之间的距离，b是决定对数螺旋形状的常数，d是在- 1和1之间随机生成的值。

3.在自然界中，鲸鱼捕猎时是绕着猎物盘旋，同时不断收缩外壳。因此，WOA被设计成模拟这种行为，假设选择收缩环绕机制和螺旋更新位置之一的概率为50%。这种行为的数学模型是:

式中:p为0 ~ 1之间随机生成的值。

在寻找猎物阶段，当收缩包围机制产生的值不在−1和1之间时，鲸鱼将放弃向最佳个体逼近，而是随机选择一个。WOA利用以下数学模型表示:

Xrand表示从当前样本中随机选择的鲸鱼。[4]

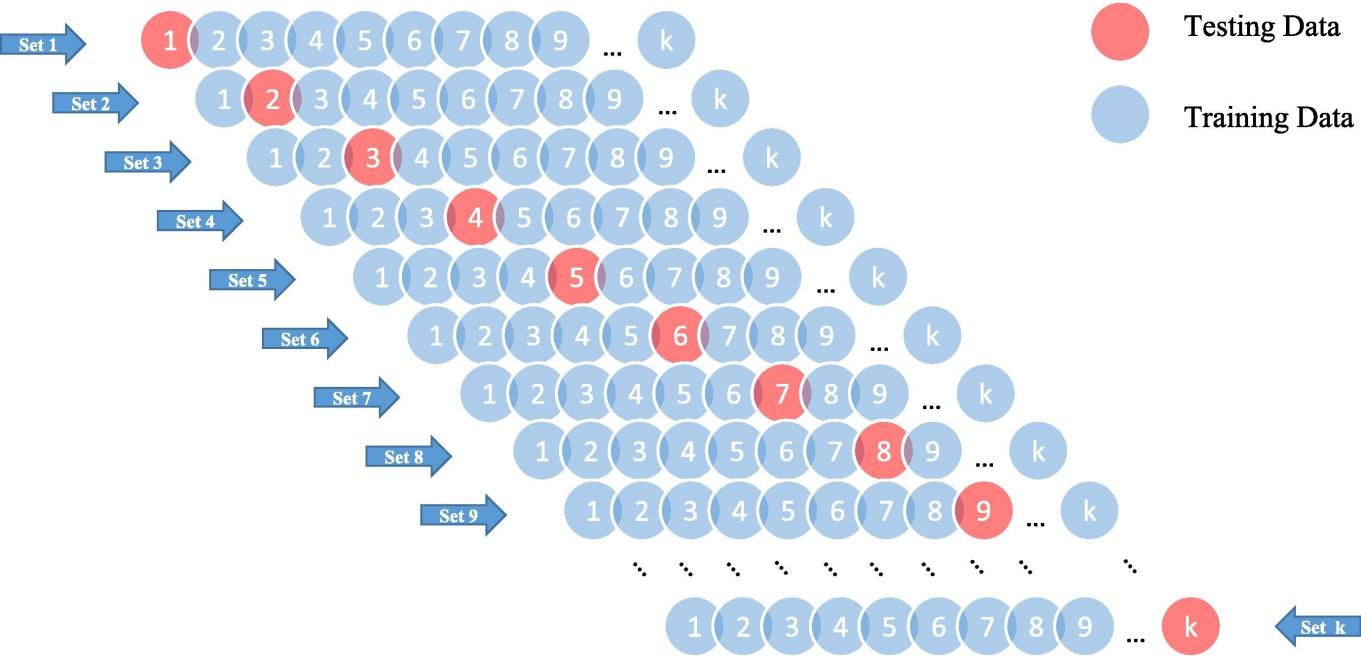
#### Fold：

K-Fold交叉验证在基于机器学习的应用中经常使用[6]。它有助于比较和选择适合特定预测分析的模型。K-Fold易于使用，用于枚举各种模型的相对效率[7]。

在本问中，使用K折交叉算法来确定参数，遵循以下过程：

(1)将收集到的训练数据集分为多个部分，分别设置为“train（1）”、“train（2）”、“train（k-1）”和“train（k）”。

(2)在交叉验证过程中，数据集的一部分被设置为测试数据集，而其余的k-1部分作为训练数据集（见图3）。平均预测精度为 MSE（均方误差），它是优化目标。



K-Fold交叉验证程序示意图[7]

十、参考文献

[1] 童明荣, 薛恒新, 林琳. 基于季节ARIMA模型的公路交通量预测. 公路交通科技. 2008(01):124-8.

[2] 吴震亚. 基于ARIMA模型的上海虹桥枢纽客流数据分析预测. 交通与运输. 2022;35(S1):315-20.

[3] Luo J, Gong Y. Air pollutant prediction based on ARIMA-WOA-LSTM model. Atmospheric Pollution Research. 2023;14(6):101761.

[4] Mirjalili S, Lewis A. The Whale Optimization Algorithm. Advances in Engineering Software. 2016;95:51-67.

[5] Nadimi-Shahraki MH, Zamani H, Asghari Varzaneh Z, Mirjalili S. A Systematic Review of the Whale Optimization Algorithm: Theoretical Foundation, Improvements, and Hybridizations. Archives of Computational Methods in Engineering. 2023;30(7):4113-59.

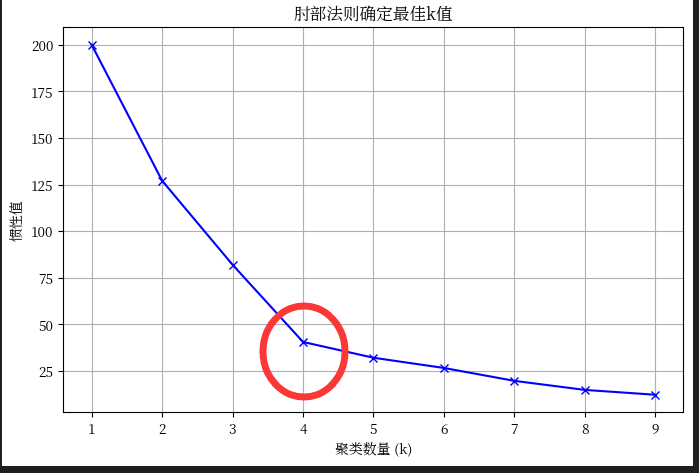
[6] Jung Y. Multiple predicting K-fold cross-validation for model selection. Journal of Nonparametric Statistics. 2018;30(1):197-215.

[7] Ling H, Qian C, Kang W, Liang C, Chen H. Combination of Support Vector Machine and K-Fold cross validation to predict compressive strength of concrete in marine environment. Construction and Building Materials. 2019;206:355-63.

### 6.2.b

#### 肘部法

“肘”方法 (Elbow method) 用于分类类别K值的确定



（1）对于n个点的数据集，迭代计算k from 1 to n，每次聚类完成后计算每个点到其所属的簇中心的距离的平方和；

（2）平方和是会逐渐变小的，直到k==n时平方和为0，因为每个点都是它所在的簇中心本身。

（3）在这个平方和变化过程中，会出现一个拐点也即“肘”点，下降率突然变缓时即认为是最佳的k值。

在决定什么时候停止训练时，肘形判据同样有效，数据通常有更多的噪音，在增加分类无法带来更多回报时，我们停止增加类别。如上图最佳分组数为4。

#### KMeans：

K均值（K-Means）：是一种常用的聚类算法，其目标是将数据集分成K个不同的簇。该算法以簇中心为基础，通过迭代更新簇中心和分配样本到最近的簇来实现聚类。计算公式如下：

(5)

式中为初始簇中心,ji是样本ji所属的簇。

**MiniBatchKMeans：**

MiniBatchKMeans：是K均值聚类的一种变体，旨在通过使用小批量（mini-batch）数据来加速和优化聚类过程，尤其适用于大规模数据集。计算公式如下：

式中，j为簇，为簇的中心。

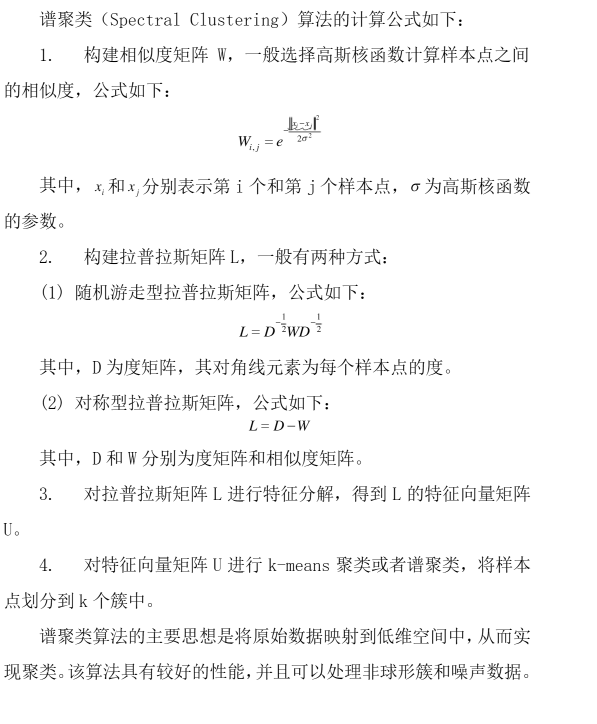
**MeanShift：**

均值漂移（MeanShift）：

式中，种子点，K是核函数，h是带宽参数，x是样本点，N(x)是以x为中心的领域。

#### SpectralClustering 公式：

原文链接：[谱聚类算法计算公式 - 百度文库 (baidu.com)](https://wenku.baidu.com/view/288ee7bca68da0116c175f0e7cd184254a351b56.html?_wkts_=1695574770753&bdQuery=SpectralClustering%E5%85%AC%E5%BC%8F)



#### AgglomerativeClustering：

凝聚层次聚类（AgglomerativeClustering）：是一种基于合并策略的层次聚类算法。它通过逐步合并最相似的簇来构建聚类结构。计算公式如下：

式中，C、Ca、Cb为簇，且C≠Ca、Cb。

#### Birch：

综合层次聚类（Birch）：BIRCH（Balanced Iterative Reducing and Clustering using Hierarchies）是一种适用于大规模数据集的层次聚类算法，旨在减少内存和计算开销。该算法通过构建一个称为CF树（Clustering Feature Tree）的数据结构来实现聚类。计算流程如下：

1. 初始化CF树：

创建一个空的CF树，并设置参数，如簇半径阈值、每个CF结点能容纳的最大样本数等。

1. 插入样本到CF树：

对于每个样本xi ：

●遍历CF树，找到最适合插入样本的叶子结点或子结点。

●如果样本能够插入到某个结点，则更新该结点的CF结构。

●如果样本无法插入到任何结点，则创建新的叶子结点，并将样本插入。

③簇分裂：

根据阈值判断是否需要对某个簇进行分裂，分裂时创建新的CF结点。

1. 簇合并：根据阈值判断是否需要对某些簇进行合并。
2. 聚类结果：遍历CF树，得到聚类结果。

BIRCH算法通过构建CF树、插入样本、簇分裂和合并等步骤来实现聚类过程。CF树结构可以高效地存储和更新聚类特征，使得BIRCH算法适用于大规模数据集，并能够生成层次化的聚类结果。

#### 轮廓系数

轮廓系数（S ）

式中a是与它同类别中其他样本的平均距离，b是与它距离最近不同类别中样本的平均距离。轮廓系数的取值范围是[-1,1]，同类别样本距离越相近不同类别样本距离越远，分数越高，分数越高说明预测越精准。

#### Calinski-Harabaz Index

方差比准则（Calinski-Harabaz Index）：

式中m为训练样本数，k是类别个数，Bk是类别之间协方差矩阵，Wk是类别内部数据协方差矩阵，tr为矩阵的迹。类别内部数据的协方差越小，组与组之间界限不明显。该指标相对于轮廓系数来说，计算速度更快，且当簇密集分离较好时，分数更高。

#### Davies-Bouldin Index

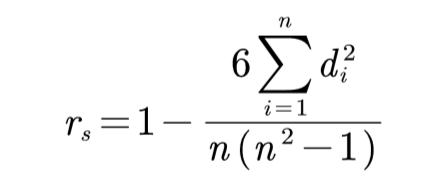
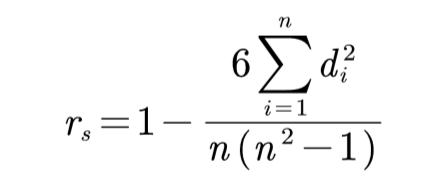
戴维森堡丁指数（Davies-Bouldin Index）：

式中n是类别个数，ci是第i个类别的中心， 是类别i中所有的点到中心的平均距离。算法生成的聚类结果越是朝着类内距离最小（类内相似性最大）和类间距离最大（类间相似性最小）变化，那么Davies-Bouldin指数就会越小。

### 6.2.c

#### 斯皮尔曼相关系数 公式

原文链接：[数学建模——相关系数（4）——斯皮尔曼相关系数（spearman）\_spearman相关系数\_Mr.Persimmon的博客-CSDN博客](https://blog.csdn.net/NoBuggie/article/details/102783973)

* 定义：X和Y为两组数据，其斯皮尔曼（等级）相关系数：
* 
* 其中，di为Xi和Yi之间的等级差。可以证明：rs位于-1和1之间。
* 定义：X和Y为两组数据，其斯皮尔曼（等级）相关系数：
* 
* 其中，di为Xi和Yi之间的等级差。可以证明：rs位于-1和1之间。

## 6.3 算法实现

### 6.3.a

### 6.3.b

### 6.3.c

### 6.3.d

### 6.4 结果分析

# 七、问题三模型的建立与求解

## 7.1 算法流程及实现

## 7.2 模型说明

### 7.2.1

### 7.2.2

### 7.3.3

## 结果分析

# 八、模型的分析与检验

## 8.1 误差分析

### 8.1.1 问题一的误差分析

### 8.1.2 问题二的误差分析

未对数据进行预处理，包括清洗，去极值，归一化等。对最终结果的准确度会造成一定影响。

### 8.1.3 问题三的误差分析

数据使用不当。

## 8.2 模型的检验

本文要解决的六个问题中模型在建立的过程中通过了相应软件的检验，具有一定的合理性。

## 8.3模型的不足

# 九、模型的评价

## 9.1模型优点

1. 算法速度快，响应性好；
2. 搜寻的结果满足所有约束要求，有较强的的实用性；
3. 建立了问题之间的联系，使得整个问题之间具备整体性；
4. 在筛选主要变量的过程中，综合了降维方法的优点，使得筛选效果更佳；

## 9.2模型缺点

问题1：

问题2：

问题3：

# 十、参考文献