|  |  |  |
| --- | --- | --- |
| 选题 | **2024年第十四届APMCM**  **亚太地区大学生数学建模竞赛（中文赛项）** | 参赛编号 |
|  | apmcm\*\*\*\*\*\* |

标题（此处换成论文的标题）

摘要

(说明：以下开始写摘要，正文从下一页开始。摘要及正文格式基本要求是宋体，小四号，单倍行距，没有要求的地方就自行处理。看完后删除该说明)

1. 问题重述

1.1问题背景

1.2问题要求

附件1、2、3给出了xxx。为了使

1. 问题分析

2.1问题一的分析

2.2问题二的分析

2.3问题三的分析

2.4问题四的分析

**文章总体思路如图1所示：**

**图1总体思路图**

1. 模型假设

四、符号说明

五、数据侧写

六、问题一模型的建立与求解

1. 问题二模型的建立与求解

7.1低，中，高风险聚类

**7.1.1.K-means++聚类过程**

K-means++是K-means聚类算法的一种改进版本，它通过智能地选择初始聚类中心来提高K-means的性能。K-means++的目标是更好地选择初始聚类中心，以减少算法的迭代次数，提高聚类的质量。具体步骤如下：

1. **选择第一个聚类中心**：随机选择一个数据点作为第一个聚类中心。
2. **选择后续聚类中心：**对于每个数据点，计算它与已选择的聚类中心之间的最短距离D(x)（即与最近的聚类中心之间的距离）。每个数据点作为下一个聚类中心的概率为:

其中x代表任意一个数据点，X表示数据集。这样保证了距离更远的数据点更有可能被选为下一个聚类中心，以确保新的聚类中心较好地覆盖数据分布。

1. **完成聚类中心的选择：**重复步骤2直到选择足够数量的聚类中心（一般为k个，我们称将数据分成了k个簇）。
2. **计算每个数据点到聚类中心的距离**：对于每个数据点，计算它与每个聚类中心的距离，通常使用欧氏距离或其他距离度量方式。对于数据点与聚类中心的距离由以下公式给出：

]

其中为第i个样本点，为第j个聚类中心，是第i个样本点的第k个分量，是第j个聚类中心的第k个分量。

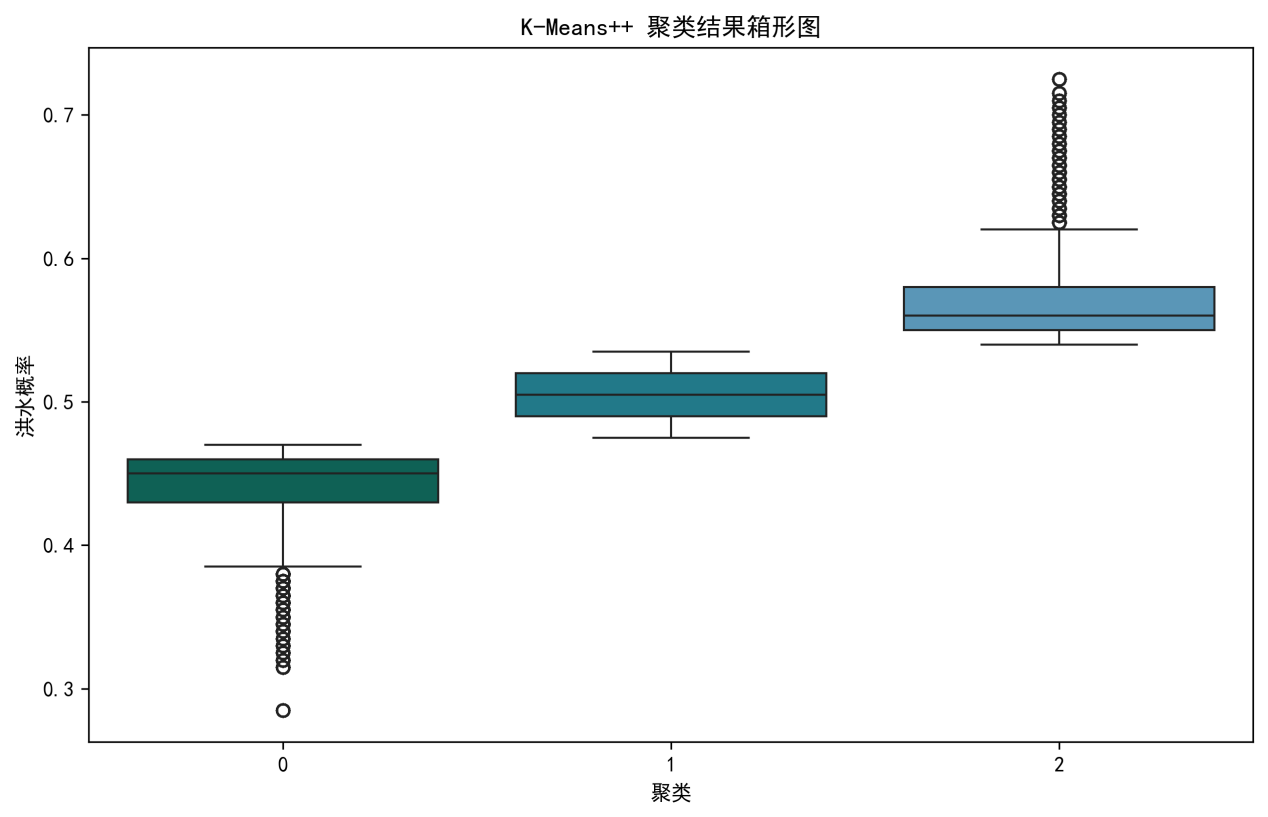
1. **分配数据点到最近的聚类中心：**将每个数据点分配到距离它最近的聚类中心所属的簇。
2. **更新聚类中心：**重新计算每个簇的中心，即该簇所有数据点的平均值。将该平均值作为新的聚类中心。即新的中心为：

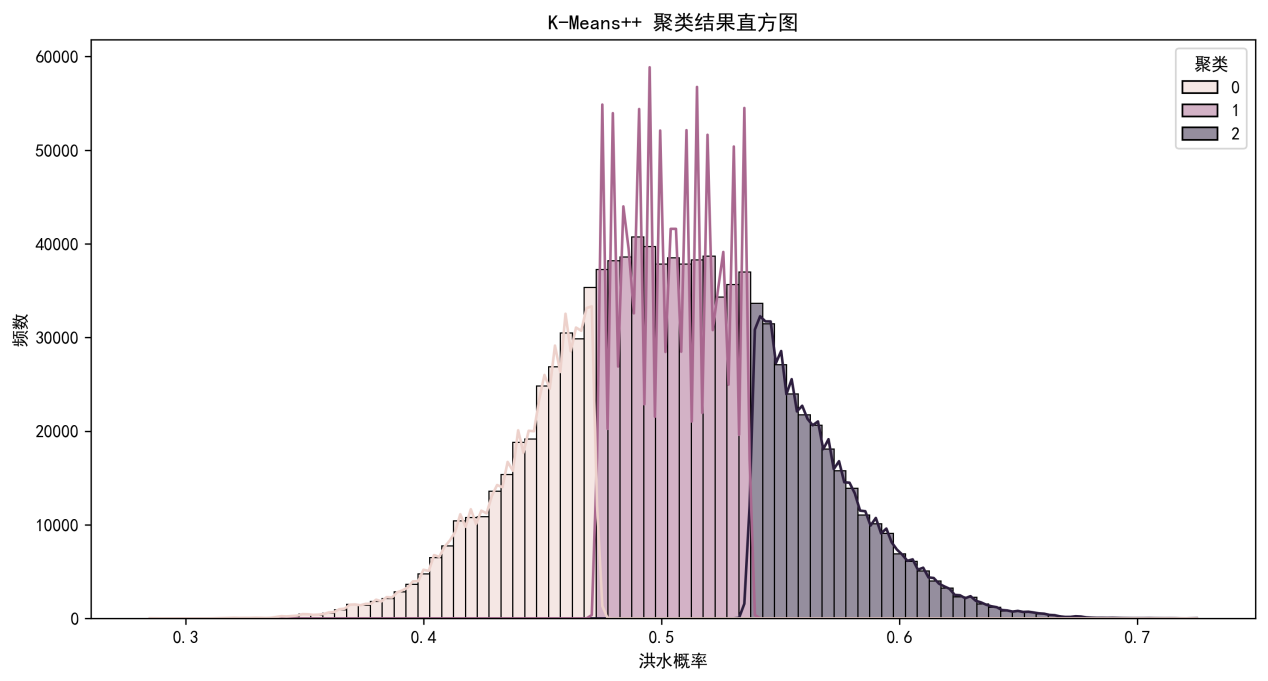
其中为新的聚类中心，其中为以为聚类中心的聚类，为中数据点的个数。

1. **完成聚类**：重复步骤4，5，6，直到聚类中心不再改变，或达到预定义的停止条件（比如达到最大迭代次数）。

最终得到每个数据点所属的簇，即完成了K-means++聚类。

使用Python进行求解，再用箱型图和直方图进行可视化，得到以下结果：





从图中不难看出，低，中，高风险的洪水概率是无交错的且大致服从正态分布，但是对于风险评级的划分，我们希望得到[0,1]区间上的三个子区间，而事实上我们K-means++得到的三个区间分别为[0.285,0.47]，[0.475,0.535]，[0.54,0.725]，显然这些区间是存在空缺的，为了填补这个空缺实现一个完整的回归区间划分使其较好地满足回归+三分类模型，基于对train.csv中洪水概率的统计分析，我们使用Logistic回归对其进行区间填补。

**7.1.2.Logistic**[**回归**](https://www.zhihu.com/search?q=%E9%80%BB%E8%BE%91%E5%9B%9E%E5%BD%92&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7b)**填补区间**

Logistic[回归](https://www.zhihu.com/search?q=%E9%80%BB%E8%BE%91%E5%9B%9E%E5%BD%92&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7b),即[对数概率回归](https://www.zhihu.com/search?q=%E5%AF%B9%E6%95%B0%E6%A6%82%E7%8E%87%E5%9B%9E%E5%BD%92&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra={"sourceType":"answer","sourceId":3526850354}" \t "https://www.zhihu.com/question/68780962/answer/_blank)，是一种线性模型，它的名字虽然叫“回归”，但却是一种用于二分类问题的算法。通过学习特征与类别之间的关系，Logistic回归可以预测新数据点属于哪个类别，输出类别的概率，其过程如下：

1. **收集数据**：收集具有标签的训练数据集，每个数据点包括特征值和类别标签。
2. **特征工程**：对数据进行特征提取和选择，包括数据清洗、特征缩放、特征选择等操作。
3. **定义预测函数**：假设我们的模型可以用如下形式表达：

其中[]是Sigmoid函数，定义为：

[]

的导数满足]。

则最终预测函数表达式为：

其中，为对应权重,需要后续过程计算得出。

最后样本为正样本的概率为，样本为负样本的概率为[。

1. **定义损失函数**：我们使用交叉熵损失函数来度量模型预测值与真实标签之间的误差（由于篇幅原因，我们不给出推导过程）：
2. **最小化损失函数：**通过梯度下降等优化算法，最小化损失函数，得到最佳的参数。
3. **预测：**当模型训练完成后，我们可以用训练好的参数对新数据进行分类预测。分类规则为：如果正样本的概率大于负样本的概率，则样本被判定为正样本，否则被判定为负样本，这等价于

于是将**7.1**中得到的三个区间两两进行Logistic回归，我们成功找到啦三个基本连续的区间[0.282500, 0.472481], [0.472481, 0.537549], [0.537549, 0.727500],事实上，这三个区间并不是严格相邻，但是，此时区间之间的距离很小，基本视为连续。因此，得出区间划分的结论为：

* **低风险区间**：**[0, 0.472481]**
* **中风险区间：[0.472481, 0.537549]**
* **高风险区间：[0.537549, 1]**

至此，我们将所有train.csv中所有数据根据洪水发生的概率划分成了低，中，高风险三个类别，接下来我们将分析三个不同类别的指标特征。

**7.2分析具有高、中、低风险的洪水事件的指标特征**

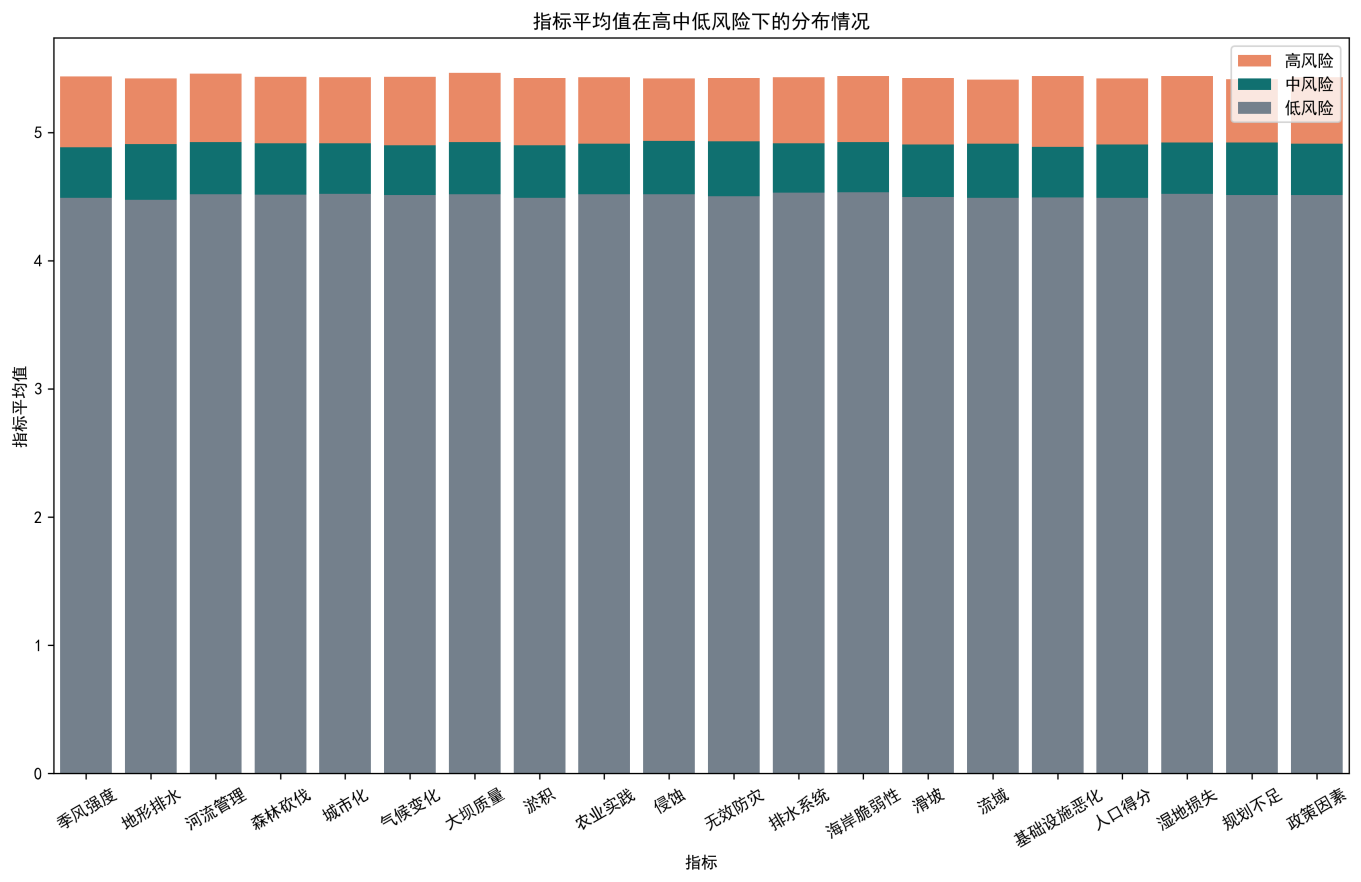
对高、中、低风险的洪水事件进行统计分析以及可视化，得如下结果（以低风险为例，其余见附表）：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 低风险区的各指标分布情况 | | | | | | | | | | |
| 指标 | 季风强度 | 地形排水 | 河流管理 | 森林砍伐 | 城市化 | 气候变化 | 大坝质量 | 淤积 | 农业实践 | 侵蚀 |
| 平均值 | 4.489 | 4.474 | 4.518 | 4.514 | 4.520 | 4.511 | 4.517 | 4.491 | 4.518 | 4.517 |
| 最小值 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 18 | 16 | 16 | 16 | 17 | 16 | 16 | 16 | 17 |
| 指标 | 无效防灾 | 排水系统 | 海岸脆弱性 | 滑坡 | 流域 | 基础设施恶化 | 人口得分 | 湿地损失 | 规划不足 | 政策因素 |
| 平均值 | 4.503 | 4.532 | 4.535 | 4.496 | 4.491 | 4.492 | 4.490 | 4.523 | 4.512 | 4.513 |
| 最小值 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 17 | 17 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 中风险区的各指标分布情况 | | | | | | | | | | |
| 指标 | 季风强度 | 地形排水 | 河流管理 | 森林砍伐 | 城市化 | 气候变化 | 大坝质量 | 淤积 | 农业实践 | 侵蚀 |
| 平均值 | 4.884 | 4.910 | 4.926 | 4.915 | 4.915 | 4.899 | 4.925 | 4.900 | 4.914 | 4.933 |
| 最小值 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 18 | 16 | 17 | 17 | 16 | 16 | 16 | 16 | 18 |
| 指标 | 无效防灾 | 排水系统 | 海岸脆弱性 | 滑坡 | 流域 | 基础设施恶化 | 人口得分 | 湿地损失 | 规划不足 | 政策因素 |
| 平均值 | 4.932 | 4.916 | 4.924 | 4.907 | 4.912 | 4.889 | 4.905 | 4.923 | 4.922 | 4.911 |
| 最小值 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 17 | 17 | 16 | 16 | 17 | 18 | 17 | 16 | 16 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 高风险区的各指标分布情况 | | | | | | | | | | |
| 指标 | 季风强度 | 地形排水 | 河流管理 | 森林砍伐 | 城市化 | 气候变化 | 大坝质量 | 淤积 | 农业实践 | 侵蚀 |
| 平均值 | 5.438 | 5.423 | 5.458 | 5.434 | 5.430 | 5.434 | 5.464 | 5.426 | 5.432 | 5.421 |
| 最小值 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 18 | 16 | 17 | 17 | 17 | 16 | 16 | 16 | 17 |
| 指标 | 无效防灾 | 排水系统 | 海岸脆弱性 | 滑坡 | 流域 | 基础设施恶化 | 人口得分 | 湿地损失 | 规划不足 | 政策因素 |
| 平均值 | 5.424 | 5.432 | 5.440 | 5.424 | 5.411 | 5.440 | 5.421 | 5.440 | 5.417 | 5.430 |
| 最小值 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 最大值 | 16 | 17 | 17 | 16 | 16 | 17 | 17 | 18 | 16 | 16 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 低，中，高风险洪水概率分布情况 | | | | | | | |
|  | 样本数 | 平均值 | 最小值 | 25% | 50% | 75% | 最大值 |
| 低风险 | 282202 | 0.4423 | 0.285 | 0.43 | 0.45 | 0.46 | 0.47 |
| 中风险 | 492938 | 0.5046 | 0.475 | 0.49 | 0.505 | 0.52 | 0.535 |
| 高风险 | 273435 | 0.5684 | 0.54 | 0.55 | 0.56 | 0.58 | 0.725 |



可以看出我们分类的结果很好地满足了分类要求，接下来我们通过线性回归建立预警评价模型。

**7.3建立发生洪水不同风险的预警评价模型**

**7.3.1线性回归构建基准模型**

线性回归是一种用于描述自变量与因变量之间线性关系的统计学方法。其基本思想是通过拟合一条直线（或超平面）来最好地拟合数据点，从而预测因变量的取值。下面详细介绍线性回归的过程和数学原理：

1. **线性关系建模**：线性回归假设因变量y与自变量x之间存在线性关系，可以用一条直线或超平面表示。数学表达式为

其中为自变量，为对应权重,[=()]称之为权重向量,[=()]称之为特征向量。

1. **最小化损失函数**：通常将均方误差（Mean Squared Error，MSE）作为损失函数，即：

其中为真实值，为预测值。

线性回归的目标是最小化损失函数，即通过最小化实际值与预测值之间的误差平方和来得到最优参数。采用最小二乘法得到最优参数

其中为一(n+1)]的矩阵，称之为特征矩阵（每一行均为不同样本点的特征向量的矩阵,故n+1为指标数+1，m为样本数目），[=()]是由对应样本的应变量构成的向量。

通过线性回归我们得到各个指标特征的权重如下：

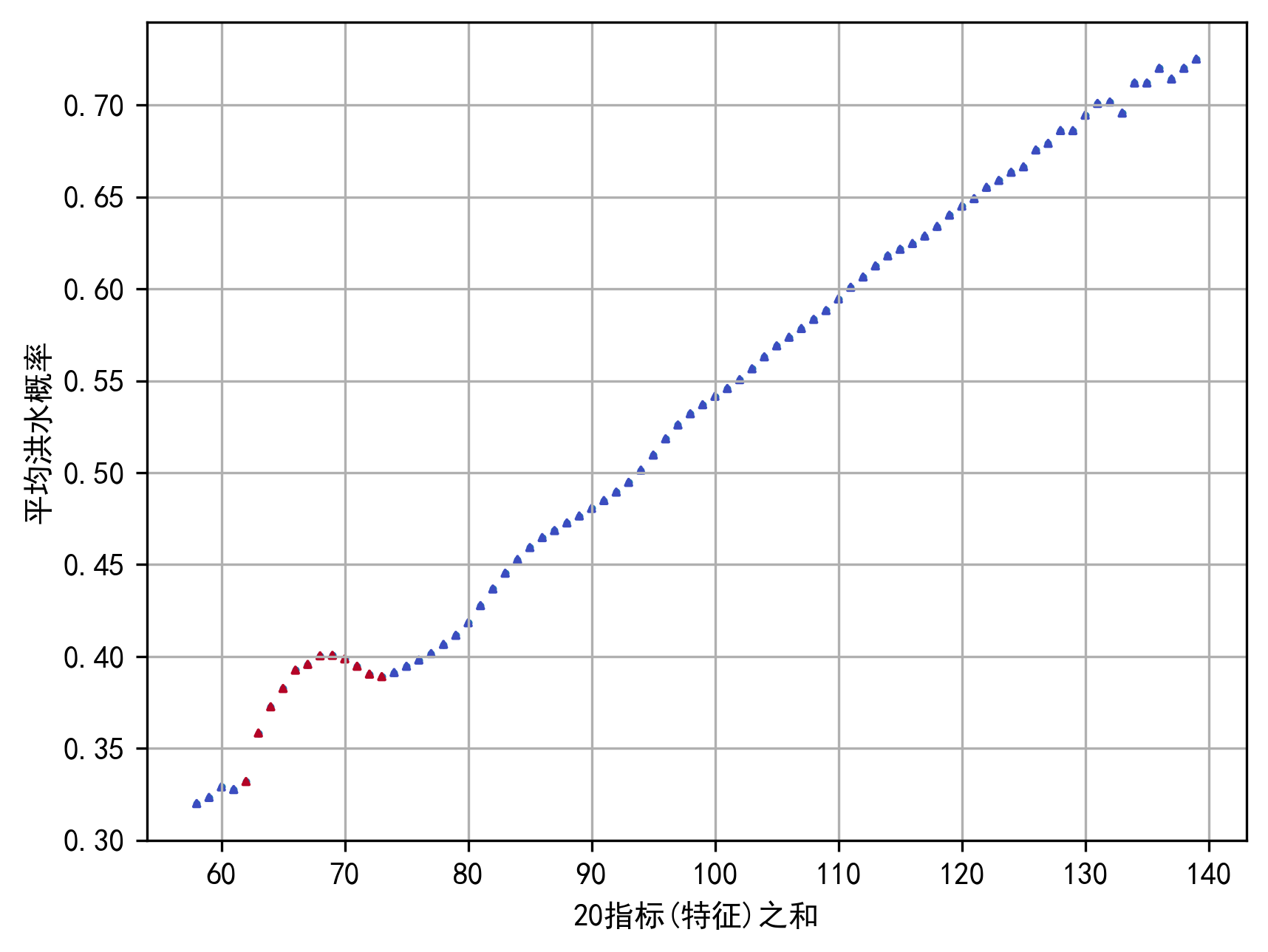
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 各指标权重（） | | | | | | | | | | |
| 指标 | 季风  强度 | 地形  排水 | 河流  管理 | 森林  砍伐 | 城市化 | 气候  变化 | 大坝  质量 | 淤积 | 农业  实践 | 侵蚀 |
| 权重 | 0.0056 | 0.0056 | 0.0057 | 0.0057 | 0.0057 | 0.0057 | 0.0057 | 0.0056 | 0.0056 | 0.0056 |
| 指标 | 无效  防灾 | 排水  系统 | 海岸脆弱性 | 滑坡 | 流域 | 基础设施恶化 | 人口  得分 | 湿地  损失 | 规划  不足 | 政策  因素 |
| 权重 | 0.0056 | 0.0056 | 0.0057 | 0.0056 | 0.0056 | 0.0056 | 0.0057 | 0.0056 | 0.0056 | 0.0056 |

至此我们得到了洪水概率预测的基准模型，接下来我们选取关键指标用以进一步建立预警评价模型。

**7.3.2选取关键指标**

在上述线性回归模型

在这一步骤，我们通过对数据的可视化探索选取**指标和**作为一个关键指标，并且将季风强度、地形排水等指标在样本维度上从小到大排序，构造**排序指标**，作为另一个关键指标（目的在于消除指标维度上的数据差异性）。



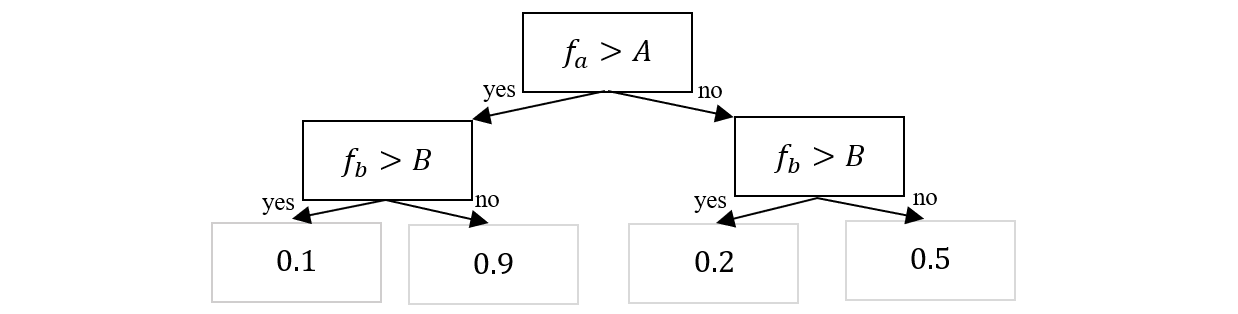
可以观察到，洪水发生概率与指标和基本呈正相关，同时，观察到线性关系中存在特殊值（红色标出部分）。不难看出，我们选取指标和作为一个关键指标，并构建排序指标是相当合理的，因为这将特殊值与平凡值都纳入了模型考虑范围。

**7.3.3CatBoost进一步建立模型**

为了更好地介绍CatBoost之前我们先简单地讲解几个概念：

* 决策树：决策树是一个预测模型，它代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。如图所示就是一个决策树。

加个公式



* **集成学习方法：**集成学习方法是一种机器学习策略，旨在通过组合多个学习器的预测结果来改善整体的预测性能。它可以通过结合多个模型的优点，来提高泛化能力和预测准确性，特别是在处理复杂问题或数据集中存在噪声的情况下表现良好。常见的集成学习方法有Bagging（Bootstrap Aggregating）， Boosting，Stacking等等，后文使用的CatBoost就是一种通过串行训练多个弱学习器，每个学习器专注于修正前序学习器的错误，最终构建一个强大的集成模型的集成学习方法。
* **梯度提升树：**梯度提升树是一种强大的集成学习方法，通过结合多棵决策树来提升预测性能。它基于逐步优化损失函数的梯度，每一棵树都专注于拟合前一棵树的残差，最终通过组合多个模型来减少预测误差，广泛应用于回归和分类任务中，特别适合处理复杂的非线性关系和高维数据。

根据以上概念我们可以得到CatBoost的基本原理。

CatBoost是一种基于对称决策树（oblivious trees）为基学习器实现的参数较少、支持类别型变量和高准确性的GBDT框架。其过程如下：

1. **定义模型：**CatBoost回归模型可以表示为一个集成模型，由多个决策树组成。假设我们有K棵树，模型可以表示为：

其中，是输入特征向量，是第k棵树的预测函数。

1. **定义损失函数：**为了训练模型，我们需要定义一个损失函数来衡量模型的预测误差。通常在回归任务中，我们使用均方误差（MSE）作为损失函数：

其中，N是训练样本的数量，是第i个样本的真实标签。

1. **训练过程:**CatBoost使用梯度提升算法来逐步优化模型。在每一步中，根据当前模型的残差计算新的决策树，使得损失函数最小化。每棵树的建立过程可以表示为：

这里 是前k-1棵树的累积预测，L是损失函数。

1. **正则化：**CatBoost在训练过程中还包括一些正则化技术，如树的深度限制、学习率控制等，以防止过拟合并提高泛化能力。
2. **预测**：训练完成后，通过将所有树的预测累加来得到最终的预测结果：

这就是CatBoost回归模型的基本数学过程：通过梯度提升的方式，逐步构建多棵决策树，并结合它们的预测来最小化损失函数，从而得到最优的回归预测模型。

为了评估我们最后得到的回归模型，我们先引入R2分数和三分类准确率。

* **R2分数：**也称为确定系数（Coefficient of Determination），是衡量回归模型拟合优度的一个常用指标,其计算公式如下：

它直观地反映了模型预测能力的强弱，并且可以用来解释模型对数据的解释程度（值越大，说明回归效果越好）。

* **三分类准确率(ACC\_3):**三分类准确率是指在一个有三个不同类别的分类问题中，模型正确预测的样本比例。通常情况下，三分类准确率可以被定义为模型对所有三个类别中正确预测的样本数目的比例,即：

这里，“预测正确的样本数”是指模型在所有预测中正确预测的样本数目，总样本数则是所有样本的数量。显然，值越大，也说明回归效果越好。

最终，我们得到的CatBoost回归模型效果如下：

|  |  |
| --- | --- |
| CatBoost回归模型效果评估 | |
| **R2分数** | **三分类准确率** |
| 0.80327 | 0.73396 |

**7.3.4 k折交叉灵敏性验证**

k折交叉验证（k-fold Cross-Validation）是一种常用的交叉验证技术，用于评估模型在数据集上的性能和泛化能力。它将数据集分成k个子集，每个子集称为一个折（fold）。k折交叉验证的过程如下：

1. **数据集划分：**

* 将数据集分成k个大致相等的部分（折）。
* 每个折依次作为验证集，其余k-1个折作为训练集。

1. **交叉验证过程：**

* 第一轮：将第一折作为验证集，其余k-1折作为训练集，训练模型并在第一折上进行评估。
* 第二轮：将第二折作为验证集，其余k-1折作为训练集，训练模型并在第二折上进行评估。
* 依此类推，直到第k轮。

1. **评估指标**：每轮验证后得到一个评估指标（比如准确率、精确率、召回率等）。

最终的模型性能评估通常是五轮验证结果的平均值，这样可以减少因为特定数据分割而引入的偏差。

此处我们采用五折交叉验证，得到的结果如下：

八、问题三模型的建立与求解

九、问题四模型的建立与求解

十、模型的评价与推广

十一、参考文献

十二、附录

**参考文献 （可另起一页）**

参考文献的编号，如[1][3]等；引用书籍还必须指出页码。参考文献按正文中的引用次序列出，其中：**书籍的表述方式为**

[编号] 作者，书名，出版地：出版社，出版年。

**参考文献中期刊杂志论文的表述方式为**

[编号] 作者，论文名，杂志名，卷期号：起止页码，出版年。

**参考文献中网上资源的表述方式为**

[编号] 作者，资源标题，网址，访问时间（年月日）。

**附录（另起一页）**