**“华为杯”第十五届中国研究生**

**数学建模竞赛**

题 目 对恐怖袭击事件记录数据的量化分析

摘 要：

恐怖袭击是指极端分子或组织人为制造的、针对但不仅限于平民及民用设施的、不符合国际人道主义的攻击行为。恐怖主义是人类的共同威胁，对恐怖事件相关数据的深入分析有助于加深人们对恐怖主义的认识，为反恐防控提供有价值的信息支持。

以下是对1998年至2017年全球恐怖袭击的分析。它具有详细的文本分析、视觉探索，以及使用各种机器学习算法例如Kmeans、决策树、随机森林、AdaBoost和bagging等模型相结合的方法进行建模，对恐怖袭击事件记录数据进行量化分析，挖掘数据集的潜在价值。

对于问题一，本文首先对恐怖袭击事件的危险等级划分进行了研究。这一研究主要分为选择重要特征对数据进行聚类和对聚类结果进行权重分级两部分。在特征提取上，由于特征值过多而且纷杂，不利于对数据进行合理的规划。为了解决这个问题，对数据中的特征进行分类提取，利用one-hot编码选择具有高方差的特征，其能够有效识别集群的特征；然后建立无监督聚类模型K-means，通过elbow方法寻找最优聚类中心个数为5，即将数据分为5类；利用主成分分析方法(PCA)计算各聚类中心的权重，可根据计算结果初步定级为一级至五级指标的权重值，通过向量乘积计算得到每类的打分值，进而初步确定为每类的危险等级，即第1类对应4级，第2类对应1级，第3类对于3级，第4类对应2级，第5类对应5级(从1级到5级的危险等级逐级递减)；最后再根据每一类的数据进行分析，确定分级的标准是否满足要求。以题一中事件编号的数据作为数据集输入模型后，即可得出相应的预测结果。

对于问题二，首先筛选2015、2016年度发生的、尚未有组织或个人宣称负责的恐怖袭击事件，然后利用无监督聚类算法K-means和层次分析法(AHP)组合成通道字典，建立模型，将可能是同一个恐怖组织或个人在不同时间、不同地点多次作案的且对应的作案组织或个人未知的若干案件归为一类，分别通过elbow方法得出最优聚类K值为5，最后经过训练集比较得出两个模型的轮廓系数分别为53.2和52.5，说明两个模型对数据集产生的效果相似，此处选择K-means方法建立最终的分级模型，建立分级模型的过程如问题一所示，其原理类似，最终分级结果为：第1类为3号嫌疑人，第2类为2号嫌疑人，第3类为5号嫌疑人，第4类为1号嫌疑人，第5类为4号嫌疑人。根据问题下表格中事件id找到对应的数据，经one-hot编码后与K-means模型得出的聚类中心进行向量相乘，得出5个代表距离的度量数据，进而可以进行1-5号嫌疑人的分级。

对于问题三，对未来反恐态势的分析，主要根据近三年(2015年-2017年)的数据来分析恐怖袭击事件发生的主要原因、时空特性、蔓延特性、级别分布等规律，如下文中的图表所示，经过分析发现，近三年内恐怖袭击最多的方式是爆炸，主要的地点分布在中东和北非地区，ISIL和Taliban是最为活跃的犯罪集团，但是总的恐怖袭击数目在逐年下降。通过分析这些特征，根据恐怖袭击在世界各地的分布及发生的频率等信息，我们建立了一个随事件索引的模型，建立了一个贝叶斯模型检验全球范围内的差异用来预测2018年的bombing/explosions的数量，进而分析研判下一年全球或某些重点地区的反恐态势，最后，根据得出的结果提出一些对反恐斗争的见解和建议。

对于问题四，如何进行数据的进一步利用，通过对1998年-2017年的恐怖袭击数据分析，结合因特网上的有关信息，我们发现伤亡人数在衡量恐怖袭击的事件严重等级中占了很大的权重，且关注生命是国际社会反恐最主要的目的，如果能准确及时的预测到此次恐怖袭击是否会有伤亡事件，能提前让当地政府和社会做好准备，可以及时的派遣足量医护人员和救援物资，避免伤亡进一步扩大。所以，本题主要是根据犯罪事件的诸多信息，预测该事件能够造成的伤亡人数，将每一类起恐怖袭击事件的死亡和受伤人数相加，对相加结果进行二分类，即伤亡数大于0设为1，否则为0，通过建立模型字典计，集合逻辑回归、支持向量机、贝叶斯、决策树、随机森林等计算多个模型，然后实例化通道将模型与训练数据相匹配，通过k折交叉验证输出每个模型在训练集上的预测精度，其中精度最高的前三个模型和预测精度分别为随机梯度下降算法：0.938，分类回归树：0.918，支持向量机：0.912.

关键词：量化分析；K-means；聚类算法；层次分析法；预测；分级

目录

1、问题背景与重述.................................................

1.1问题背景..................................................

1.2问题的提出................................................

2、模型假设与符号说明.............................................

2.1模型假设..................................................

2.2 符号假设.................................................

3、问题一模型建立与求解..........................................

3.1问题描述及分析............................................

3.2 模型建立..................................................

3.2问题求解与分析..............................................

4、问题二模型建立与求解............................................

4.1问题描述及分析..............................................

4.2模型建立....................................................

4.3问题求解与分析..............................................

5、问题三模型建立与求解.............................................

5.1问题描述及分析...............................................

5.2模型建立.....................................................

5.3问题求解与分析...............................................

6、问题四模型建立与求解.............................................

6.1问题描述及分析...............................................

6.2模型建立.....................................................

6.3问题求解与分析..............................................

1. 模型的评价与改进建议............................................

7.1模型评价..................................................

7.2 模型与方法的改进建议......................................

参考文献..........................................................

附录...............................................................

# 问题背景与重述

## 问题背景

近年来，全球的经贸、科技等都在向着更加发展的方向进行，但是，世界各国都避不开的一个话题就是恐怖袭击。目前，恐怖袭击事件在全球分布很广，也是全球地缘政治问题日益加剧具体体现。此外，GTD关于1998-2017年的恐怖袭击事件统计数据显示，近20年来，已经从1998年的935件年发生数量达到了2017年的10897件，上升了9倍有余。其中，更是在2014年达到了16902件，为历史之最。

普通人虽然也可以通过互联网、电视等媒体接收到相关信息，各个国家对于相关事件也都有记录。但是，随着信息技术的发展，人们希望对恐怖袭击有更深的了解，包括但不限于发生地、产生因素以及恐怖组织等基本信息。

更深层次的是，期望对于全球各恐怖袭击事件之间的关系能够有更清晰的认识。首先，就是对恐怖袭击事件危险程度的认识。然而，国际上关于恐怖袭击的**危险等级**并没有一个统一的衡量标准，因此，我们希望能够通过对数据分析，通过对数据进行**量化处理**，建立一个标准，以求能够对恐怖事件进行定级，更进一步希望能够对犯罪嫌疑人进行筛选。同时，能够完成对于恐怖事件处理方面的相关预测行为。

## 问题提出

问题一：

# 模型假设与符号说明

## 模型假设

(1)、假设数据独立同分布；

(2)、假设默认选择的参数如聚类中心有很好的拟合性；

(3)、假设部分数据编码后符合正太分布；

(4)、假设题种的高斯过程,数据是联合高斯分布的；

(5)、假设贝叶斯分类过程中符合独立条件性假设。

## 符号说明

表2-1 变量符号说明

|  |  |
| --- | --- |
| 符号 | 说明 |
| x | 标量 |
| D | 数据样本（数据集） |
| （·，·，） | 行向量 |
| （·；·；） | 列向量 |
| {…..} | 集合 |
| I | 单位阵 |
| H | 假设集 |
| |{…..}| | 集合{….}中元素个数 |

## 4 问题一

## 3.1 问题描述及分析

本题是根据附件1中的数据，考虑人员伤亡和经济损失，及与发生的时机、地域、针对的对象等等诸多因素，结合现代信息处理技术，借助数学建模方法建立基于数据分析的量化分级模型，根据建立的模型将附件1给出的事件按危害程度从高到低分为一至五级，并且需要列出近二十年来危害程度最高的十大恐怖袭击事件，同时根据表1中给出的事件id对该事件进行分级。

本题给出的数据中，从1998年到2017年世界各地的恐怖袭击事件，总共有114184条，包括国家、地区、伤亡人数、攻击类型等135个特征，其中包括了60000多起爆炸事件、7000多起暗杀事件和8000起绑架案等。而每个案例至少有45个变量信息，最近的事件包括超过120个变量的信息。针对如此多的特征变量，就不能使用通常的主观方法进行分级，我们希望计算机能够自动处理数据并分为不同的类别，由于数据集中没有目标变量，所以这是一个典型的无监督学习问题，即根据类别未知的训练样本解决模式识别中的各种问题。对于这类分级问题，首先想到的是聚类方法，无监督学习的聚类方法分为两类，一类是划分聚类方法，即通过优化评价函数把数据集分割为K个部分。典型的划分聚类算法有K-means算法, K-medoids算法、CLARANS算法。层次聚类由不同层次的分割聚类组成，层次之间的分割具有嵌套的关系。典型的分层聚类算法有BIRCH算法、DBSCAN算法等。由于K-means算法简单、快速，对于处理大数据集，该算法能保持可伸缩性和高效性。因此本题选择K-means聚类算法。

通过K-means算法可以得到簇类质心，本题要求需要将危害程度分为五级，所以选取的K值必须大于等于5，当模型最优K值大于5时，需要将合并质心，将分类限定为五级；当K值等于5时，类别即可认为是级数，此时需要将类别排名，即每一类所代表的事件的危害程度。利用PCA方法，计算权重，将权重与每个质心向量相乘，得到5组数据，可以看作是每种类别危害程度。通过分析每类的事件，验证分级结果是否满足标准，即可得到分级类型，然后将表中的事件放入模型中，得到相应分级的结果。

## 模型建立

解决本题的分级问题需要建立两个模型，一个是K-means聚类算法，能够将附件1中的数据分为五类；二是PCA主成分分析法，确定聚类结果每类的权重，用来对聚类结果进行分级。

下图为模型建立的流程图：

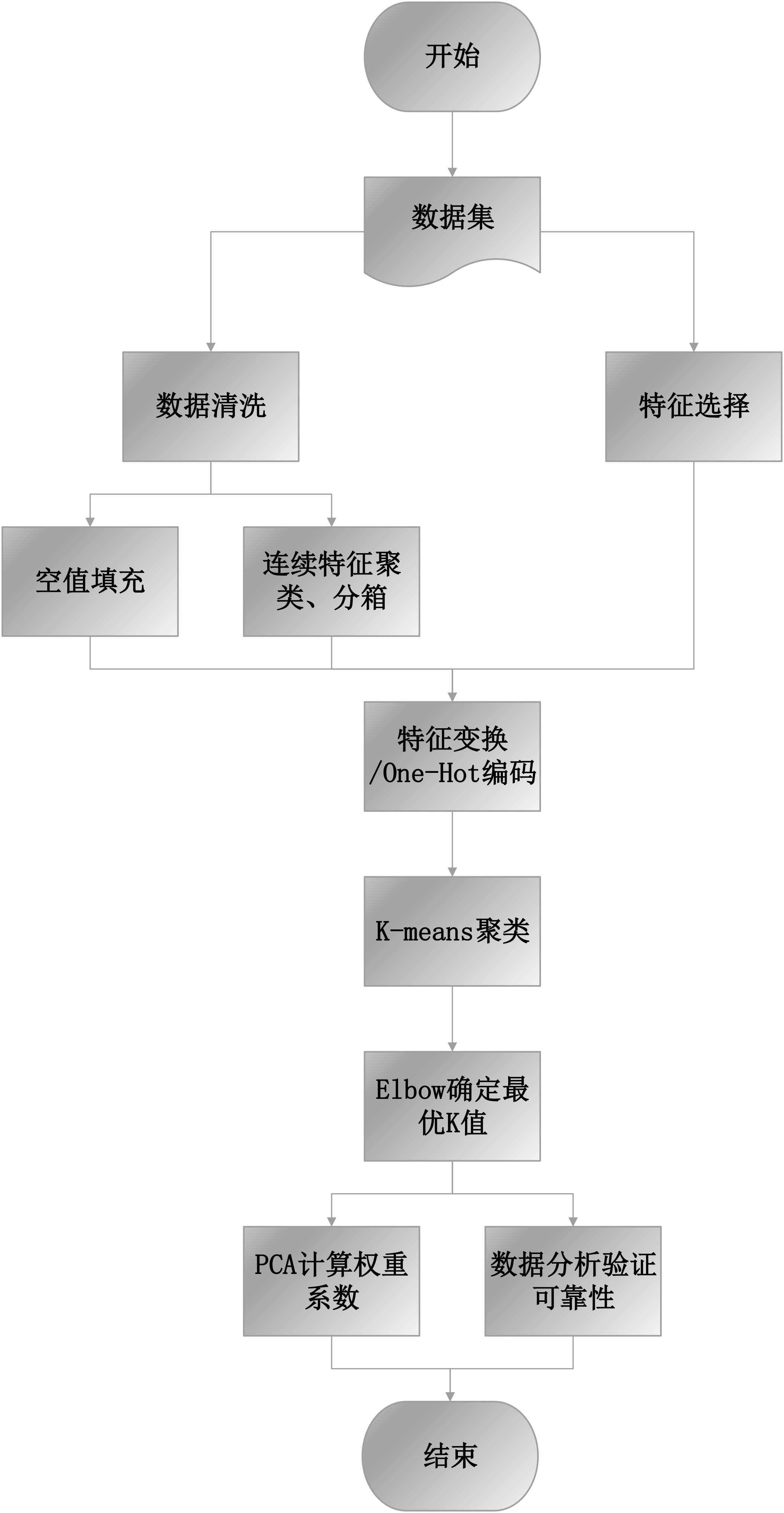


图3-1：模型流程图

首先加载附件1中的数据，在建立模型之前需要对数据进行预处理，包括数据清洗和特征选择，填补异常值和空值，对于’nkill’列和’nwound’列，采用中值填充缺失值，同时为了避免噪声点的影响，将连续类型’nkill’和’nwound’列的数据分箱，本题为了拟合数据本身的特点，利用K-means将’nkill’和’nwound’列分成了六类，分类结果分布图如图3-2所示，从图中可以看出，分类满足要求。

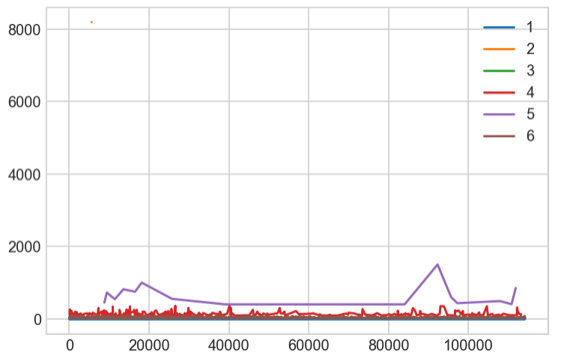
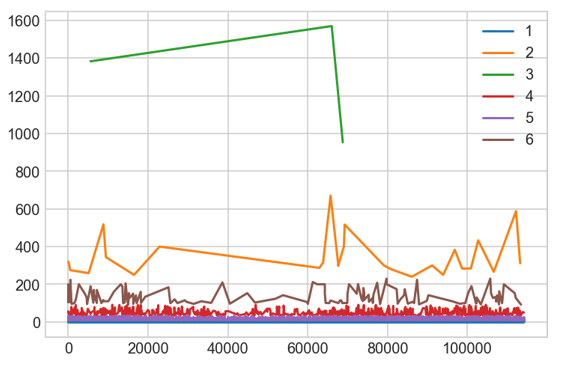


图3-2 ‘nkill’列(左)和’nwound’列(右)分箱结果分布图

本题选择国家、地区、攻击类型、伤亡人数、财产等17个特征作为聚类标准，然后所选的特征进行one-hot编码，并选择具有足够高方差的变量，这些变量可以更好的识别集群，经过预处理后，数据变成61维的数据。此时的数据可作为训练数据放入模型进行训练。我们通过寻找高方差方法确定并选区了61个关键特征指标作为关键标签，并通过K-means算法设计了多级指标表单，利用PCA方法给予对应的初始权重值。

## 模型一：K-means聚类算法

K-means算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。首先要确立K个初始质心，对于要进行聚类的每一个对象，计算与K个质心的距离并把其分布到距离最近的聚类质心的簇中，然后下一个对象重复计算这个过程，直到所有质心都不再发生变化，即所有的簇都收敛。

k-means算法是一种很常见的聚类算法，它的基本思想是：通过迭代寻找k个聚类的一种划分方案，使得用这k个聚类的均值来代表相应各类样本时所得的总体误差最小。

k-means算法的基础是最小误差平方和准则。其代价函数是：

式中，表示第i个聚类的均值。我们希望代价函数最小，直观的来说，各类内的样本越相似，其与该类均值间的误差平方越小，对所有类所得到的误差平方求和，即可验证分为k类时，各聚类是否是最优的。

上式的代价函数无法用解析的方法最小化，只能有迭代的方法。k-means算法是将样本聚类成 k个簇（cluster），其中k是用户给定的，其求解过程非常直观简单。

聚类划分用于对于一个包含n个多维对象的集合D，划分出K(K<=n)个子集合，每个子集合就是一个簇，K-means算法最关键的问题是寻找最优K值，本题通过elbow方法循环查找，最后根据轮廓系数的大小，得出最优子集合为5个。

算法具体实现过程如下：

(1) 从数据集中选取K个对象作为质心(K限定在2-11的区间之间)；

(2) 根据每个簇的质心，计算剩余点到质心的距离，形成新的簇；

(3) 重新计算已经得到的各个类的质心；

(4) 迭代2～3步直至新的质心与原质心相等或小于指定阈值，算法结束；

(5) 根据输入的训练集计算模型的轮廓系数，重复开始执行第(1)步；直到K值到达区间的最右端点；

(6) 此时分别得到K取2~10时的轮廓系数，比较这9个轮廓系数的大小，取最优的K值，经计算取得K为5时聚类效果最好，此时轮廓系数约为60%。

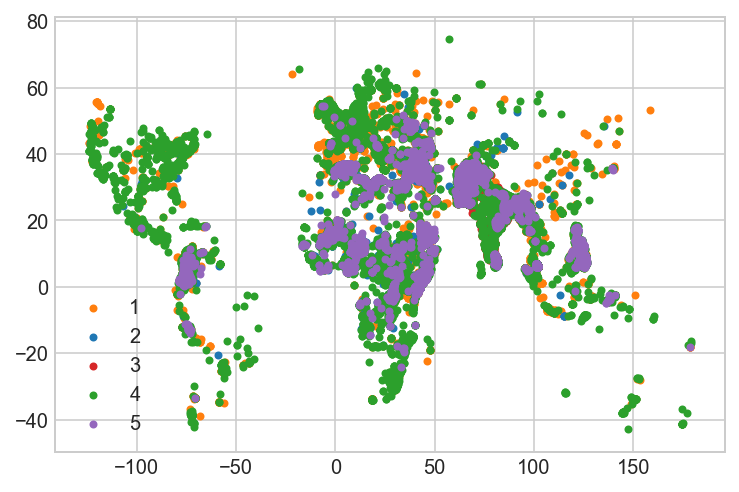
利用K-means算法获得的分类结果及对应的数目如下表所示：

|  |  |
| --- | --- |
| 类别 | 数目 |
| 3 | 37268 |
| 4 | 33157 |
| 5 | 19199 |
| 2 | 14376 |
| 1 | 10183 |

K-means算法获得的分类结果对应的具体事件分布情况如下表所示，主观上可以判断类别2的危害性最大，与PCA计算结果相符：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **类别**  **指标** | **1** | **2** | **3** | **4** | **5** |
| 死亡人数标准差 | 5.999 | 37.492 | 1.519 | 5.416 | 6.035 |
| 受伤人数标准差 | 12.018 | 144.508 | 5.944 | 5.248 | 4.010 |
| 财产损失重大以上的个数所占的比例: | 0.784% | 1.571% | 0.344% | 0.371% | 0.046% |
| 常用武器类型 | Explosives | Explosives | Explosives | Firearms | Explosives |
| 常用攻击类型 | Bombing/  Explosion | Bombing/  Explosion | Bombing/  Explosion | Armed Assault | Bombing/  Explosion |
| 目标/受害者 | Private  Citizens  &  Property | Private  Citizens  &  Property | Private  Citizens  &  Property | Private  Citizens  &  Property | Military |
| 最多的受害人群 | Civilians | Civilians | Civilians | Civilians | Soldiers |
| 最活跃的犯罪集团 | ISIL | ISIL | Taliban | Taliban | AI-Shabaab |

K-means聚类结果地区分布图如下图所示：从图中可知，聚类事件多分布在中东和南非一带。



## 3.2.2 模型二：主成分分析（PCA）算法确定权重

主成分分析法通过变量变换的方法把相关的变量转换为若干不相关的变量，主要用来对数据降维，在本题中，主要用来确定权重系数。主成分分析的一般模型为：

式中，，，…，为分析后得到的m个主成分；为决策矩阵中系数。

通过主成分分析确定权重主要在专家打分的基础上，其整个过程类似上述公式，对与本题中的部分专家模型可表示为下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 专家 指标 | 专家1 | 专家2 | 专家3 | 专家4 | 专家5 | 专家6 |
| 死亡人数 | 5 | 5 | 4 | 5 | 5 | 5 |
| 财产损失 | 4 | 5 | 3 | 5 | 4 | 4 |
| 持续时间 | 4 | 4 | 4 | 3 | 4 | 3 |
| 武器类型 | 3 | 2 | 2 | 2 | 3 | 2 |
| 受害目标 | 5 | 5 | 5 | 4 | 5 | 5 |

每个专家所打分的权重不同，且存在一定的偏向，从而给权重的确定带来一定的模糊性。根据专家打分的思路，PCA确定权重的主要步骤为：

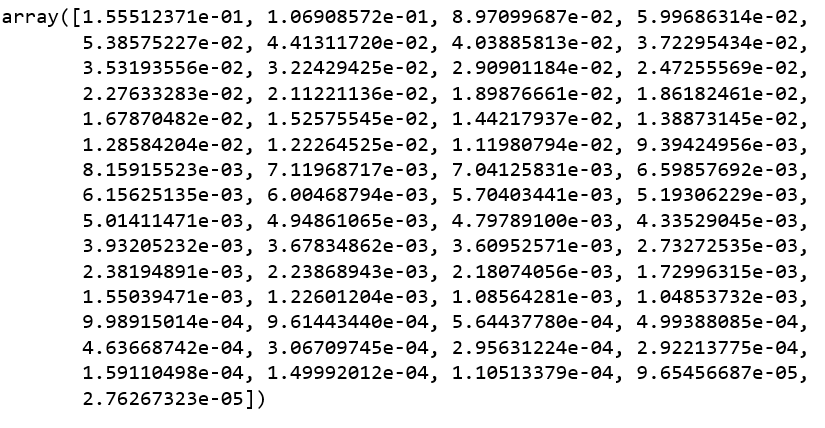
指标权重等于以主成分的方差贡献率为权重，对该指标在各主成分线性组合中的系数的加权平均的归一化，因此，要确定指标权重需要知道三点：

A. 指标在各主成分线性组合中的系数；

B. 主成分的方差贡献率;

C. 指标权重的归一化.

本题通过Sklearn的PCA函数计算训练集的权重系数，其中特征个数选取为训练数集的列数，计算得到的权重系数如下图所示：



PCA计算的权重向量

然后，将计算得到的权重向量与K-means聚类算法得到的聚类中心向量相乘，即得到如下表所示的每个类别对于的分数，该分数根据大小值的不同可初步认定分数越大其分级级别就越高。

|  |  |
| --- | --- |
| 类别 | 分数 |
| 1 | 0.368 |
| 2 | 0.496 |
| 3 | 0.377 |
| 4 | 0.391 |
| 5 | 0.363 |

## 问题求解与分析

问题1：将附件1给出的事件按危害程度从高到低分为一至五级

依照上诉问题分析和模型建立的过程，解决本题的方法是1、建立聚类模型，将附件1的事件分为五类；2、通过PCA方法确定权重，并对每类事件进行分析验证，根据评分得到每一类的级数。

其中建模前的数据预处理过程、特征的选择过程和建模过程如上所述。

最后得出类别与危害程度等级的对应标准如下表所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类别 | 1 | 2 | 3 | 4 | 5 |
| 危害程度 | 4 | 1 | 3 | 2 | 5 |

问题2：二十年来危害程度最高的十大恐怖袭击事件

根据上述，通过PCA主成分分析方法确定了权重，利用选取危险等级为一级的事件，与权重系数分别向量相乘，得到每个事件的评分，按照评分从高到底排列，选取Top10.

二十年来危害程度最高的十大恐怖袭击事件如下表所示：

|  |  |
| --- | --- |
| 事件编号 | 危害程度(从高到低) |
| 200109110004 | 1 |
| 200403110001 | 2 |
| 200409010002 | 3 |
| 200507070001 | 4 |
| 200811260003 | 5 |
| 201101240016 | 6 |
| 201107220011 | 7 |
| 201403010012 | 8 |
| 199808070002 | 9 |
| 200210230004 | 10 |

问题3：对表1的事件进行分级

首先根据事件编号与附件1中的’eventid’进行匹配，获得该事件的全部数据，按照建模过程数据预处理的方法，先对数据进行清洗，将清洗后的数据代入K-means训练好的模型中，得出该事件所属类别编号，然后根据类别—危险等级标准表，查找该事件所属的危害等级。

表1的事件分级如下表所示：

表1 典型事件危害等级

|  |  |
| --- | --- |
| 事件编号 | 危害等级 |
| 200108110012 | 二级 |
| 200511180002 | 一级 |
| 200901170021 | 二级 |
| 201402110015 | 一级 |
| 201405010071 | 一级 |
| 201411070002 | 三级 |
| 201412160041 | 一级 |
| 201508010015 | 三级 |
| 201705080012 | 一级 |

## 问题改进

本题在建模的过程中是使用了K-means一种方法，无法证明对于本问题来说该模型是最优的模型，因此，应该尝试基于密度的聚类算法(如：DBSCAN)、层次分析方法等，分别尝试训练模型，根据评价函数选取最优方案。

利用PCA方法进行计算权重的过程中，无法避免评分结果与实际不符，为提高准确率，对于特征选择采取更多样的方法，尝试添加新的特征或减少特征能否对模型的准确率有所提升。

# 问题二

## 问题描述及分析

将附件1中时间在2015、2016年度发生的、尚未有组织或个人宣称负责的恐怖袭击，且作案组织或个人未知的事件。运用数学建模方法，将可能是同一个恐怖组织或个人在不同时间、不同地点多次作案的若干案件归为一类，并将对应的未知作案组织或个人标记不同的代号，按该组织或个人的危害性从大到小选出其中的前5个，记为1号-5号。再对表2列出的恐袭事件，按嫌疑程度对5个嫌疑人排序，并将结果填入下表2中。

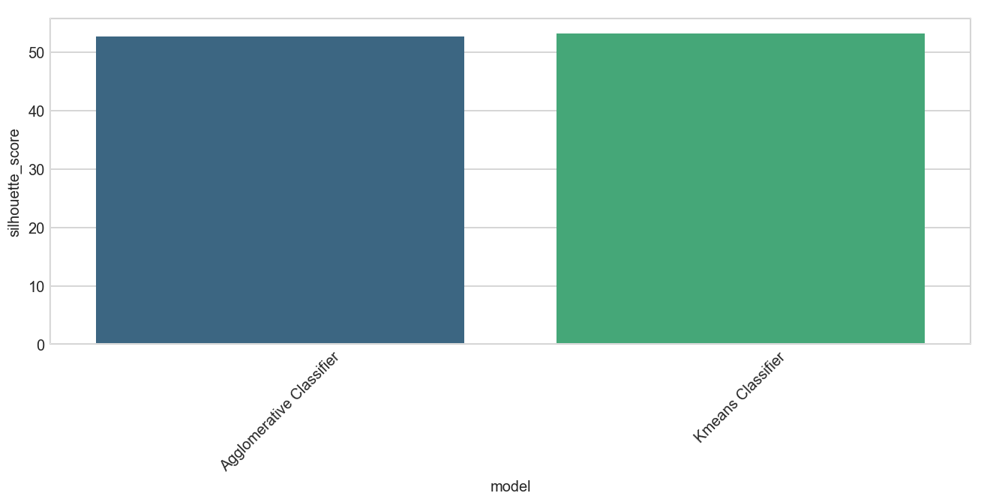
本题与上一题类似，也是分类模型。本题的数据集进一步收缩，事件限制在2015年和2016年发生的，且尚未由组织或个人宣称负责(‘claimed’字段限制为1)的恐怖袭击(‘crit1’、’crit2’和’crit3’字段分别限制为1,且’doubtter’字段限制为0)，作案组织或个人未知(‘gname’ 字段限制为 'Unknown')的事件，对该事件集进行聚类，把相似的事件归为同一类别，本题聚类的个数依然限制在5个以上。然后按照组织或个人的危害性从大到小排序，选出前5个，类似与问题一中的分级模型。

最后，需要对表2中的恐怖袭击事件，按照嫌疑程度对5个嫌疑人排序。解决此问题的方法与问题一中的方法类似，首先根据聚类模型训练数据集，得出k(k>=5)个聚类中心，然后将表2中的事件编号与数据集相匹配，获得该事件的全量数据，然后将经过数据预处理的该事件的数据与聚类中心向量相乘，此时获得k个数，按照获得的k个数大小进行排序，最后根据其对应的聚类类别查找到对应的危害性。

## 模型建立

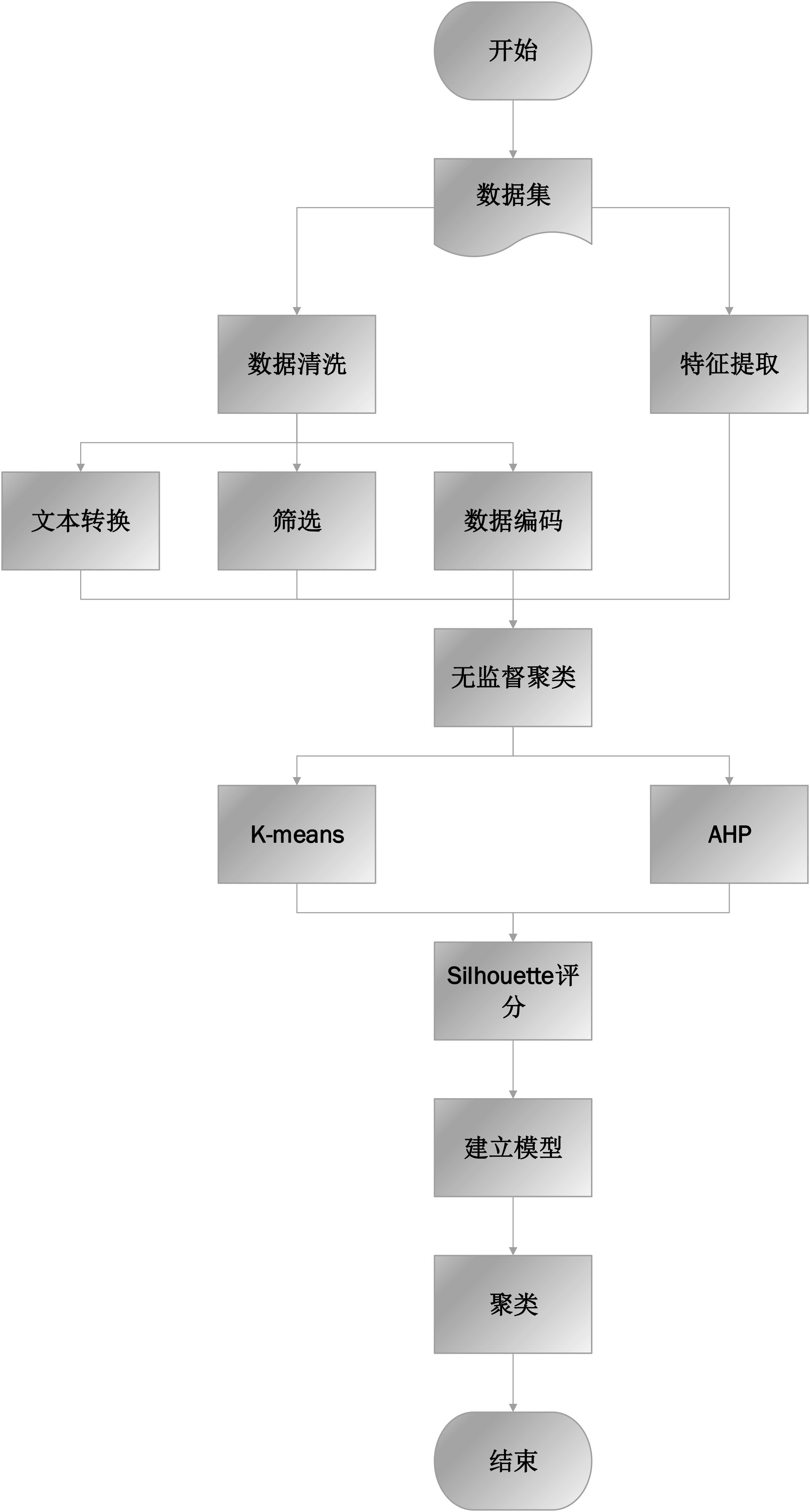
解决本题的问题需要建立聚类模型和评分模型，在问题一建立分级模型的时候，我们使用的是K-means算法，其效果达到了我们想要的需求，在本题中，利用K-means算法和层次聚类算法构造模型字典，然后实例化通道将模型与训练数据相匹配，利用silhouette轮廓系数评价标准，选择评价最高的算法用来建模。

下图为K-means算法和层次聚类算法在训练集上通过轮廓系数评价的结果，可以看出，K-means算法和层次聚类算法对于该训练集的拟合效果相似，不过，本题还是选择了评价效果更好一些的K-means算法训练模型，其具体聚类建模的过程如问题一所述，此处不再赘述，本题主要介绍层次聚类算法。



评分模型依然采用问题一所示的PCA主成分分析法和专家评分相结合的形式。

下图为问题二模型建立的流程图：



模型流程图

4.2.1 模型一：层次聚类算法

层次聚类可在不同层次上对数据集进行划分，形成树状的聚类结构。AggregativeClustering是一种常用的层次聚类算法。

假设有N个待聚类的样本，对于层次聚类来说，基本步骤就是：

(1) 把每个样本归为一类，计算每两个类之间的距离，也就是样本与样本之间的相似度；

(2) 寻找各个类之间最近的两个类，把他们归为一类（这样类的总数就少了一个）；

(3) 重新计算新生成的这个类与各个旧类之间的相似度；

(4) 重复(2)和(3)直到所有样本点都归为一类，结束

其原理是：最初将每个对象看成一个簇，然后将这些簇根据某种规则被一步步合并，就这样不断合并直到达到预设的簇类个数，由于每个簇就是一个集合，因此需要计算出集合之间的距离，常用的距离计算方法有三个，分别是：两个簇的样本对之间距离的最小值、两个簇的样本对之间距离的最大值和两个簇的样本对之间距离的平均值。

本题中采用的层次聚类模型，指定簇类的个数5，链接算法采用最小值来度量。

## 问题分析与求解

问题1：将2015、2016年度发生的、尚未有组织或个人宣称负责的恐怖袭击，且作案组织或个人未知的事件。运用数学建模方法，将可能是同一个恐怖组织或个人在不同时间、不同地点多次作案的若干案件归为一类。

本题采用与问题一类似的聚类算法，即通过聚类算法将互为相似的事件聚类一类。建模前的数据预处理过程与上诉的方式一样，需要选择‘claimed’、‘crit1’、’crit2’和’crit3’字段分别为1,且’doubtter’字段为0，’gname’字段为 'Unknown'的子数据集，然后进行聚类算法模型的建立，本题采用K-means和层次聚类算法组合构造字典模型， 然后实例化通道将模型与训练数据相匹配，利用silhouette轮廓系数评价标准，选择评价最高的算法用来建模，根据评价便准的结果，选择K-means算法建模，其相对于训练集的聚类结果如下表所示：

|  |  |
| --- | --- |
| 类别 | 数目 |
| 1 | 5845 |
| 5 | 3679 |
| 2 | 1654 |
| 3 | 1005 |
| 4 | 126 |

问题2：按照组织或个人的危害性从大到小排序选择其中的前5个，记为1号-5号。

本题也与问题一中的分级相似，对于训练好的模型，通过PCA主成分分析方法计算每个簇类向量的权重，然后将权重向量与簇类向量相乘，得到k个值(本题k值为5)，根据这k个值的大小，可以初步认定每类簇所代表的危害程度。最后通过数据分析方法，分析每一类的数据特点，综合评价并验证PCA计算权重系数的准确度。

下表为权重向量与簇类中心向量的乘积，可以看出第4类的得分最高：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类别 | 1 | 2 | 3 | 4 | 5 |
| 得分 | 59.15 | 70.11 | 41.29 | 72.19 | 55.37 |

根据每一类的具体分析，得出簇类别与危险性的对应表如下所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类别 | 1 | 2 | 3 | 4 | 5 |
| 危险性 | 3号 | 2号 | 5号 | 1号 | 4号 |

问题3：对表4-1列出的恐袭事件，按嫌疑程度对5个嫌疑人排序，填表。

通过表4-1中的事件编号与附件1中的数据相匹配，得到该 事件对应的全量数据；然后通过建模前的数据预处理阶段，对该事件的数据进行清洗，然后将清洗过的数据与训练模型得到的训练簇类中心向量(此处有5个)相乘，分别计算出5个不同的数值，根据这5个值排序，最后根据该数值对应的聚类类别在查找其所属的危险等级，即可按照这些值排序出对应的嫌疑人。

表4-1 恐怖分子关于典型事件的嫌疑度

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1号嫌疑人 | 2号嫌疑人 | 3号嫌疑人 | 4号嫌疑人 | 5号嫌疑人 |
| 201701090031 | 5 | 4 | 3 | 2 | 1 |
| 201702210037 | 5 | 4 | 3 | 2 | 1 |
| 201703120023 | 4 | 5 | 3 | 2 | 1 |
| 201705050009 | 5 | 4 | 2 | 3 | 1 |
| 201705050010 | 5 | 3 | 4 | 2 | 1 |
| 201707010028 | 5 | 4 | 3 | 1 | 2 |
| 201707020006 | 5 | 4 | 3 | 2 | 1 |
| 201708110018 | 5 | 3 | 4 | 2 | 1 |
| 201711010006 | 5 | 3 | 4 | 2 | 1 |
| 201712010003 | 5 | 4 | 3 | 1 | 2 |

## 问题改进

本题在建模的过程中是采用K-means和分层聚类算法构造字典模型，根据轮廓系数的评分标准，选择最优算法，对于本题来说是一种优化改进，但是由于特征系数过多，且问题最后要对五人进行排序，所以最后的误差可能会有所差距，应该结合多种算法进行组合模型，且使用多种评分方式。

# 问题三

## 问题描述与分析

为了未来提高反恐的效率和提高针对性，依据附件1的数据并集合伊特网上的有关信息，建立适当的数学模型，集合图表，研究三年内，即2015年-2017年的恐怖袭击事件发生的主要原因、时空特性、蔓延特性、级别分布等规律，进而能够研究判断下一年全球或某些重点地区的反恐态势。

本题的数据集用到的是附近1中的数据，且需要把’iyear’字段限制在2015年-2017年之间，且事件是恐怖袭击的事件，根据数据可视化图标，根据时空、地域、级别、原因等不同角度分许近三年恐怖袭击事件的规律。通过了解近三年的数据分布情况，对恐怖袭击事件随时间的趋势有一个大概的了解，进而可以预测下一年的恐怖袭击情况，通过建立相关的数学模型，比如逻辑回归、贝叶斯等。

## 模型建立

本题采用贝叶斯算法，朴素贝叶斯分类器建立在“贝叶斯假设”的基础上：给定目标值时属性之间相互条件独立，根据经典的贝叶斯公式：来对时间进行预测。

恐怖袭击是贝叶斯推理研究的一个成熟领域。考虑到它们的频率，很难假设大量的样本接近正态分布。正因为如此，应该预先建立一个已知区域所看到的恐怖程度，并在新的信息（如的一年年的袭击或来自同一地区的对比国家）之前更新。

使用贝叶斯推理来比较选择的两种类型。进而需要确定的是，一个地区的恐怖程度是否与另一个地区（或时间周期）有显著的不同。

例如，如果要知道南美洲的一个国家是否与另一个地区有很大的不同，可能会预先假定一些国家是南美的一个国家，在南美国家中平均受到攻击和变异。然后，会在南美的一个国家和南美洲的一个独立国家的信息中更新这一信息。

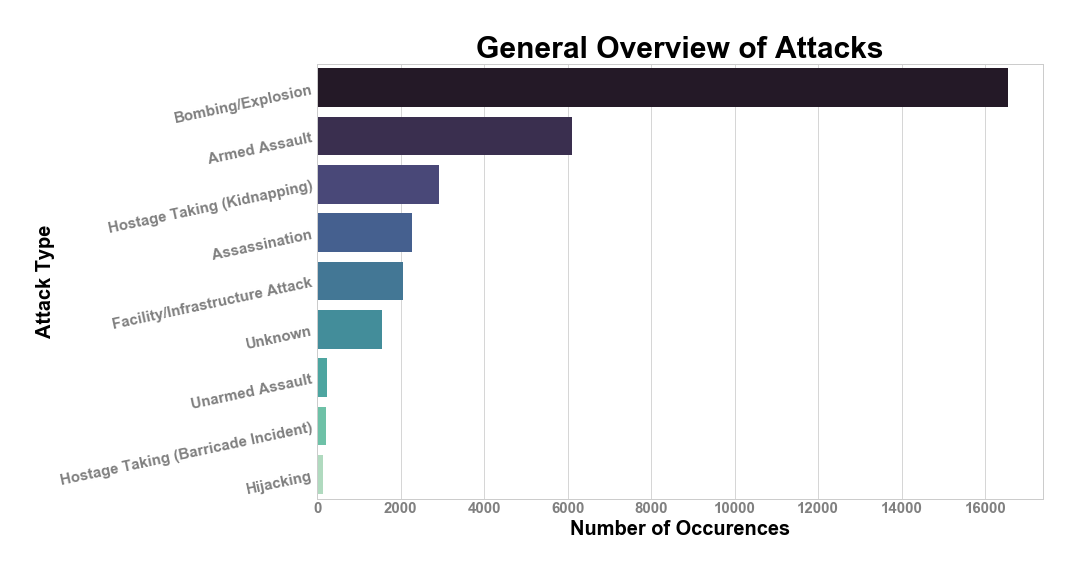
本题中，利用贝叶斯深度网络pymc3构建一个贝叶斯模型，根据Iraq 和 Columbia的活动预测2018年发生的bombing/explosions的数量。首先在数据集中找到凶手数量的平均值和方差，分别为：13.37、87.82，然后将数据带入训练模型，通过极大后验估计 MAP,使用最优化方法找到点估计。

## 问题求解与分析

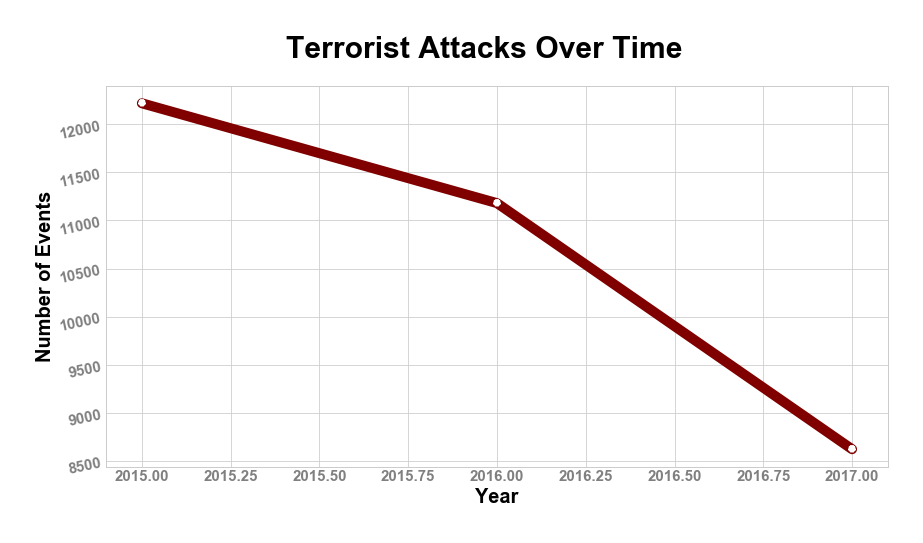
本题的思路主要包括探索性数据分析，然后贝叶斯推理来检验全球范围内的差异。

首先，应该了解恐怖袭击攻击的方式、在世界各地的分布，以及它们的频率，需要生成一个被时间索引额可视化图表。

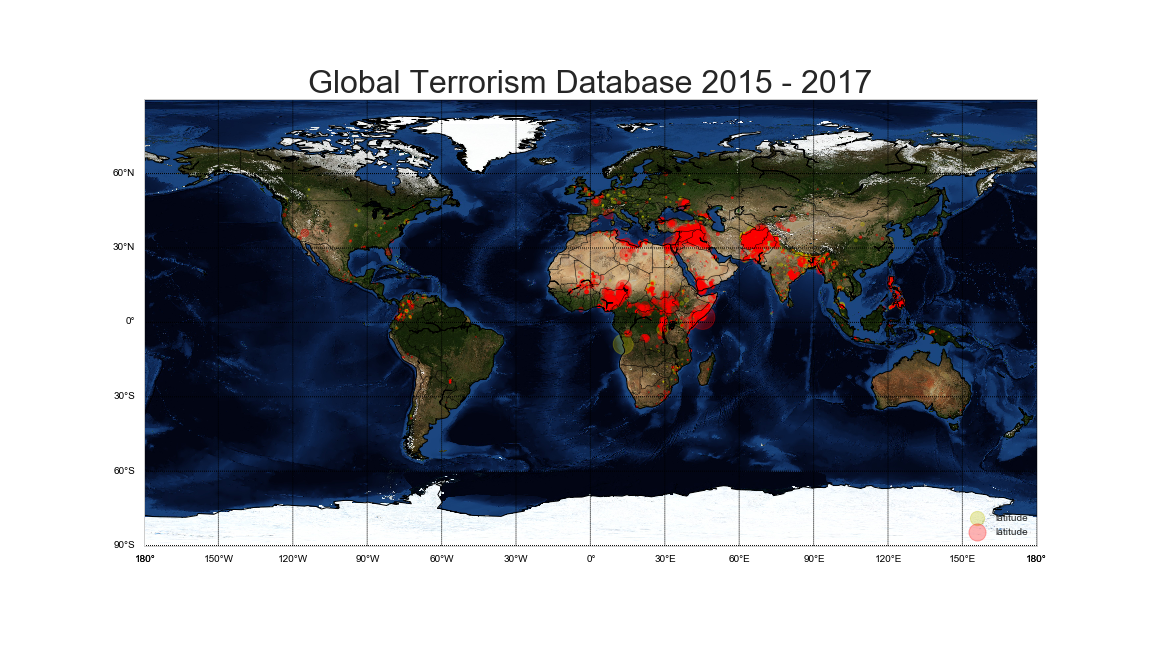
如图所示，显示了2015年-2017年恐怖袭击攻击方式的分类，爆炸式袭击占据超过所有攻击方式一半的比重。



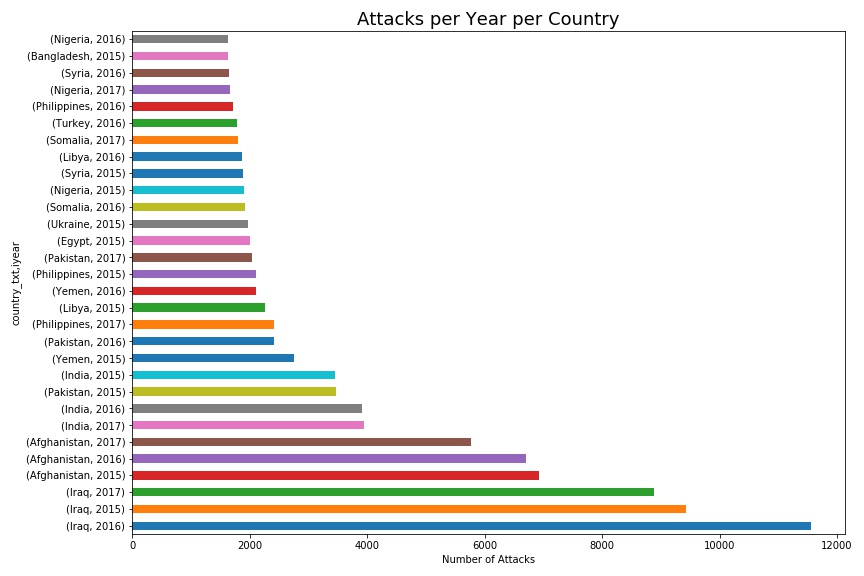
随着时间的推移，每年的攻击数量在下降



根据2015年-2017年恐怖袭击事件地区分布图可知，恐怖袭击事件主要聚集在非洲、澳大利亚、东亚等地区。



根据每个国家每年的恐怖袭击数来看，属于战争地区的伊拉克遥遥领先。



根据以上的分析结果，我们可以建立相应的模型预测未来一年的恐怖袭击趋势，本文主要以预测发生爆炸事件，通过上述的分析结果，建立深度贝叶斯模型，通过第三方深度贝叶斯库pymc3，可以很方便的建立贝叶斯模型。

## 问题改进

本题主要是针对某一方便的事件进行预测，比如本题代码部分通过贝叶斯预测爆炸情况，但是对于诸如伤亡、财产损失等事件不适用。考虑创建管道来拟合各种情况。

.

# 问题四

## 问题描述与分析

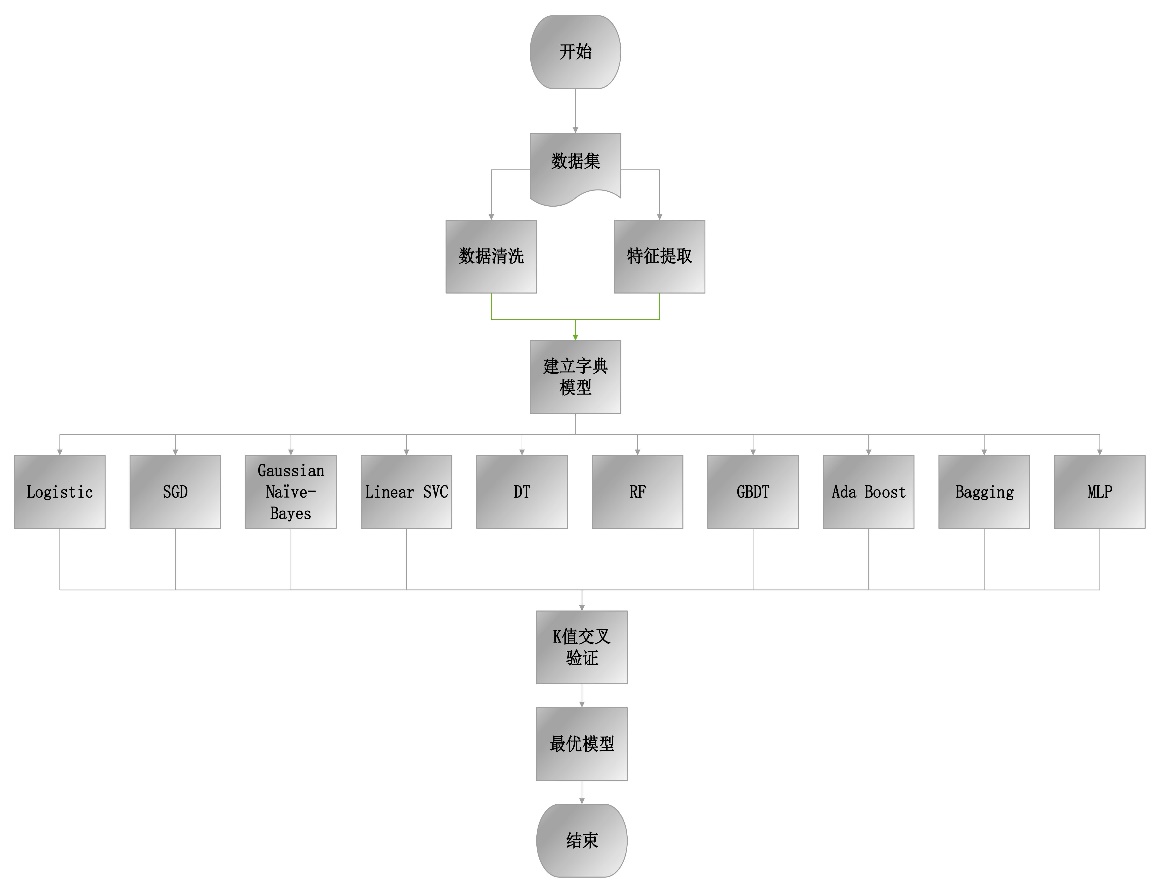
本题主要是对数据的进一步使用，具有发散性思维，由于附件1中的数据较大，且特征较多，所能挖掘的价值很多，通过对1998年-2017年的恐怖袭击数据分析，结合因特网上的有关信息，我们发现伤亡人数在衡量恐怖袭击的事件严重等级中占了很大的权重，且关注生命是国际社会反恐最主要的目的，如果能准确及时的预测到此次恐怖袭击是否会有伤亡事件，能提前让当地政府和社会做好准备，比如医疗人员随时等候等，避免伤亡进一步扩大。

所以，本题的解决方案主要是根据犯罪事件的诸多信息，预测该事件是否会造成伤亡。首先将附件1中的数据受伤人数和死亡人数相加，得到单独的一列内容，为一个事件死亡和受伤的总人数。通过将伤亡人数大于0的设为1，否则设为0的方法，可以将建模方法变为分类模型，大大简化了建模的难度和复杂度。

## 模型建立

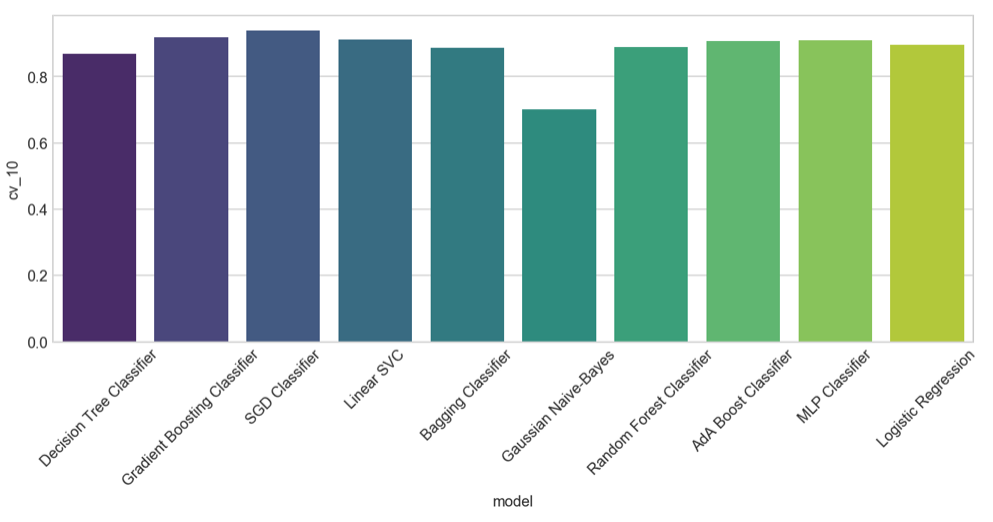
本题被简化为一个二分类模型，有多种算法可以来实现这个任务，我们想要尽可能的把更多、更全面的模型都跑一遍，进而比较各种模型在训练集上的集合程度，从中选择最优的模型。

如建模流程图所示，经过最初的出局预处理之后，通过建立一个字典模型，集合逻辑回归、支持向量机、贝叶斯分类器、随机森林等多个模型，然后实例化通道将模型与训练数据相匹配，通过k折交叉验证方法输出每个模型在训练集上的预测精度，根据预测精度的大小选择合适的分类器用于创建最终的模型。



## 问题分析与求解

本题是一个发散性思维题，有多个价值能够挖掘，本题的思路主要是预测未来发生的恐怖袭击是否会造成伤亡，而不是去预测具体的伤亡人数。正因为如此，我们正努力使模型的TP最大化，减少我们的FN，从而努力使模型的查全率最大化。如上所述，我们在建立模型的时候，通过字典创建了多个模型来测试哪个模型对于训练集的效果最佳，经过k折交叉验证方法，计算出每个模型在训练集上的精确度即查全率，如下图所示：



从图中可以看出，除了NB，模型整体的拟合程度很相似，在训练集上表现最好的前五个模型为：SGD、GBDT、Linear SVC、AdaBoost和LR。

基于树的模型在这里用于预测，因为它们对异常值更健壮，而且由于数据有许多异常值，尽管有了数据集，基于树的模型是我们最好的选择，这是因为，像逻辑回归或神经网络这样的模型对异常值非常敏感，而基于树的模型并不是这样的，因为它们根本不考虑数据点有多远。它们所看到的只是通过分割一个属性来获得“信息增益”（或减少数据的随机性），这与它对目标变量的分类有更多的关系。

此外，基于树的模型是快速学习算法和数据集也是巨大的，基于如表所示的各算法的优缺点，本题在最终建模的过程中选择了随机森林这一算法，在于能够找到特征的重要性评分，可以平衡误差，避免过拟合。通过随机森林构建的预测模型，在测试集上的准确率为 89.52%。

|  |  |  |
| --- | --- | --- |
| 算法 | 优点 | 缺点 |
| Logistic | 实现简单，计算速度快 | 容易欠拟合，分类精度不高 |
| SGD | 样本训练速度快，可在线更新 | 更新不一定按正确方向进行，会带来扰动 |
| SVC | 解决非线性分类效果好 | 大规模数据训练比较苦难，无法直接支持多分类 |
| Gaussian | 小规模数据表现很好，适合多酚类和增量训练 | 对输入数据的表达形式敏感，分类性能不一定高 |
| DT | 对缺失值不敏感，对异常点容错能力好，不需要提前归一化，可处理多维度输出分类问题 | 容易过拟合，只能注重单个特征的局部划分 |
| RF | 并行处理速度快，可很好避免过拟合，能够找到特征的重要性评分，可以平衡误差 | 偏差会增大 |
| GB | 弱分类，能自我纠错 | 过拟合 |
| AdaBoost | 具有较低泛化误差，具有更高精度，不容易出现过拟合现象 | 迭代次数不太好设定，数据不平衡导致分类精度下降，数据训练耗时 |
| Bagging | 避免过拟合，训练样本之间独立 | 预测函数只能顺序生成，训练耗时 |
| MLP | 高度并行性，高度非线性全局作用，良好的容错性，十分强的自适应和自学习能力 | 网络中隐含节点个数选取困难，学习速度慢，容易陷入局部极值 |

## 问题改进

特征的过多加重了模型的冗余度，多个模型相结合出现的训练结果，无法对单一算法进行单独调优。应该在代码中添加模块，可以分别调整模型的参数。

# 模型评价

## 模型的优点

对于问题一：采用K-means聚类与PCA计算权重系数相结合的方法，对事件进行分级。为避免离群点的干扰，在数据预处理阶段，对于连续型变量采用了K均值方法对该变量进行分类。

对于问题二：此题与问题一类似，在聚类建模阶段，采用K-means聚类与层次聚类相结合的模式，选择在训练集上拟合程度更好的模型。

对于问题三：采用贝叶斯深度网络，在时序关系上建立预测模型。

对于问题四：此题是发散性思维的题目，自主挖掘数据的潜在价值。本文将回归问题简化为二分类问题，且组合多个类型的算法进行预测，交叉验证选择最优的结果。

## 模型的缺点

对于问题一：采用单一的K-means聚类算法，该算法对离群点敏感，由于数据集中有大量离群点，所以会对模型有一定的影响。

对于问题二：由于数据集特征系数过多，且问题最后要对五人进行排序，利用与聚类中心向量相乘的方法判断五个指标，跟实际的情况可能会有所差距。

对于问题三：模型只能针对特定的情况，跟时间密切想关，鲁棒性不好。

对于问题四：特征的过多加重了模型的冗余度，多个模型相结合出现的训练结果，无法对单一算法进行单独调优。