** **

**“华为杯”第十五届中国研究生**

**数学建模竞赛**

|  |  |
| --- | --- |
| **学 校** | **温州大学** |
| **参赛队号** | **18103510007** |
| **队员姓名** | 1. **姜贺云** |
| **2.刘新伟** |
| **3.张振宇** |

**“华为杯”第十五届中国研究生**

**数学建模竞赛**

题 目 C

摘 要：

和平与发展是当今时代的主题，恐怖主义日益成为国际政治中的重要话语和国际安全领域的重大问题，对世界形势和格局的影响也是十分复杂和深远的。但对于反恐防恐的意识还待加强，本文立足于某组织搜集整理的全球恐怖主义数据库，通过四个问题的分析和研究，采用聚类分析模糊评价恐怖袭击事件的危害等级，随机森林分析潜在的嫌疑人以及相关数理统计知识，为反恐防恐提供有价值的信息支持。

对于问题一，先对数据进行预处理，选取本文仅仅选择缺失值比例小于50%的特征来减少对研究任务的影响，最后确定8个特征，采用K-means聚类方法，得出5个随机聚类块，之后对属性设置权重，这里采用熵法确定权重，依据上述属性的权重确定值，计算随机聚成5类的概率，概率大者危害等级较高，则可得出危害等级一至五级，以及十大恐怖事件（见表1）。

对于问题二，对2015、2016犯罪集团（每一个集团的攻击次数）数据集分为测试集和验证集，测试集按照70%和30%比例分为测试集和验证集，之后采用随机森林分类算法，对数据处理，选取重要的前50个特征，然后进行预测，其中排名前5的组织为：Muslim extremists，Taliban，Maoists，Separatists，Fulani extremists。案件进行分类预测，分类结果达94%，说明模型具有合理性，然后利用Softmax多分类回归模型，对每个样例输出一个概率值，选出前5个对应的概率值进而判断犯罪嫌疑人（见表3），发现新生或者隐藏的恐怖分子，尽早采取相应的措施，防微杜渐。

对于问题三，我们应用数理统计的知识，先对日期进行数据处理，然后分析随着时间的变化，全球以及某些地区的反恐态势，得出恐怖袭击事件在前期是以缓慢增长的态势，在中期有个极具下降的趋势，在之后，恐怖袭击案件是以指数增长趋势发展，恐怖袭击案件最高发地区为东欧，中东南非，西欧，亚洲南，从2010开始，各地恐怖袭击事件发生的次数明显比之前多。

对于问题四，应用数理统计知识，分析随着时间的变化攻击手段和次数以及攻击目标和使用的武器，得出轰炸，爆炸 是犯罪分子最常使用的犯罪武器，经常攻击的目标是商业，军队，警方和普通公民。

本文是基于大数据的处理，分析恐怖事件的危害性，潜在性以及防范措施等，这有利于我国有效的推进“一带一路”战略，也有利于各国在经济政治文化全球化背景下实现双赢。

关键词：大数据处理，恐怖主义，熵方法，Softmax多分类回归，随机森林，数理统计，反恐措施

目录

[一、问题重述](#_Toc514104625) 3

[二、问题分析](#_Toc514104626) 3

[三、模型假设](#_Toc514104627) 4

[四、定义与符号说明](#_Toc514104628)

[五、模型的建立与求解](#_Toc514104629)

[第一部分：准备工作](#_Toc514104630)

[第二部分：问题1数学模型](#_Toc514104631)

问题2数学模型…………………………………………………………………………………………….

[问题3数学模型](#_Toc514104632)

问题4数学模型 …………………………………………………………………………………………

[六、模型评价与推广](#_Toc514104633)

[七、参考文献](#_Toc514104634)

[八、附件](#_Toc514104635)

一 问题重述

由于恐怖袭击事件的危害性不仅取决于人员伤亡和经济损失这两个方面，还与发生的时机、地域、针对的对象等等诸多因素有关，因此对附件1中给出的事件建立量化分析，按危害程度从高到低分为一至五级，列出近二十年来危害程度最高的十大恐怖袭击事件，并给出表1中事件的分级。

根据附件1中的多起恐怖袭击事件尚未确定作案者进行研究，针对在2015、2016年度发生的、尚未有组织或个人宣称负责的恐怖袭击事件，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串联起来统一组织侦査，寻找可能性，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，对应的未知作案组织或个人标记不同的代号，并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。再对列出的恐袭事件，按嫌疑程度对5个嫌疑人排序。

对未来反恐态势的分析评估有助于提高反恐斗争的针对性和效率。依据附件1并结合网上的有关信息，研究近三年来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，进而分析研判下一年全球或某些重点地区的反恐态势，用图/表给出研究结果，提出对反恐斗争的见解和建议。

利用附件1的数据研究恐怖袭击主要攻击的手段类型，主要攻击目标和使用的武器。这样在应对恐怖袭击时，知己知彼，做好相应的安保措施和预防手段。

二 问题分析

2.1问题一

首先，对数据进行预处理，选取本文仅仅选择缺失值比例小于50%的特征来减少对研究任务的影响。在此，记缺失值比例小于50%的特征组成的数据矩阵特征矩阵。

对事件进行划分危害等级前，先找到数值变量集合，本文取数值变量集合与特征矩阵的特征列集合的交集作为可计算的数值变量，得到可以用来做计算的特征总共有8个，该数据集的样本数中存在着缺失值，我们对缺失值采用均值的填充方法。确定分级条件，采用聚类方法得出危害等级。

2.2问题二

对2015、2016犯罪集团（每一个集团的攻击次数）数据集分为测试集和验证集，测试集按照70%和30%比例分为测试集和验证集，之后采用随机森林分类算法，对案件进行分类预测，分类结果达92%，说明模型合理性，进而判断犯罪嫌疑人。通过分类算法的研究，这样有助于提高破案效率，有利于尽早发现新生或者隐藏的恐怖分子，尽早采取相应的措施，防微杜渐。

2.3问题三

我们应用数理统计的知识，先对日期进行数据处理，然后分析随着时间的变化，全球以及某些地区的反恐态势，得出恐怖袭击事件在前期是以缓慢增长的态势，在中期有个极具下降的趋势，在之后，恐怖袭击案件是以指数增长趋势发展，恐怖袭击案件最高发地区为东欧，中东南非，西欧，亚洲南，从2010开始，各地恐怖袭击事件发生的次数明显比之前多。

2.4问题四

应用数理统计知识，然后分析随着时间的变化攻击手段和次数以及攻击目标和使用的武器，得出轰炸，爆炸 是犯罪分子最常使用的犯罪武器，经常攻击的目标是商业，军队，警方和普通公民。

三、模型假设

1．假设题目所给的数据真实可靠；

2．假设数据各个特征是相互独立的；

3.

四、定义与符号说明

|  |
| --- |
| M ——表示特征数目 |
| N——训练样本的个数 |
| m ——表示输入特征数目 |
| ——分类结果 |
| ——V个分支节点D所在属性a上的取值 |
| V——表示分支节点 |
| pi——表示事件xi发生的概率。 |

五、模型的建立与求解

第一部分：

数据的处理

1. 本文舍弃缺失值超过90%的特征列，通过计算得到剩下的特征列占总特征的比例为53%，剩下的特征中缺失值比例在50%-90%仅仅占其19%，因此，本文选取仅仅选择缺失值比例小于50%的特征来减少对研究任务的影响。在此，记缺失值比例小于50%的特征组成的数据矩阵特征矩阵。
2. 通过处理可得到58个特征（详见附件1），其中17个文本特征，27个分类变量，8个计算特征，6个额外特征（详见附录2）；
3. 数据残缺的问题，可以根据数据挖掘等理论根据变化趋势进行补充，本文采用对8个可计算特征缺失值做均值处理；
4. 对17个文本特征和27个分类变量，缺失值和unknow记为-1处理。

恐怖袭击定义：

　　国际上对于恐怖主义犯罪的界定有很多种不同的说法，综合比较熟知的十几种定义来说，可以把恐怖主义犯罪归纳为以下几种：

第一类认为对人类的生命造成威胁或者带来恐怖的一切暴力性质的活动都是恐怖主义犯罪，强调的是恐怖主义的暴力性及恐怖性，比较严重的刑事犯罪都是恐怖主义犯罪，比如杀人、防火、绑架等。

第二类认为有预谋或者政治目的的一些恐怖活动都是恐怖犯罪，强调的是具有预谋性和政治性的恐怖主义活动，这一定义是国际上认同度比较高的概念。例如，英国的反恐法案中指出“出于政治目的而使用暴力，致使公共场合混乱公众恐怖的活动”。这一概念一般指，恐怖主义是有预谋，有计划的。

　尽管国际上对于恐怖主义犯罪的概念不同，但是恐怖主义犯罪都有某些特定的特点是相同的。如：恐怖活动的动机越来越多，并且越来越多样化，恐怖活动的形式呈现多样化。恐怖活动的主体呈现出多样化和小型化的特点，因为种族、宗教等问题在世界范围内一直都存在，此外，恐怖主义还有鲜明的政治性，恐怖活动往往是有鲜明的政治目的的[1]。

第二部分：

问题1的数学模型

对数据进行预处理，选取本文仅仅选择缺失值比例小于50%的特征来减少对研究任务的影响。在此，记缺失值比例小于50%的特征组成的数据矩阵特征矩阵。 对事件进行划分危害等级前，先找到数值变量集合，本文取数值变量集合与特征矩阵的特征列集合的交集作为可计算的数值变量，得到可以用来做计算的特征总共有8个：美国死亡人数（nkillus），凶手死亡人数(nkillte)，抓获的凶手数量(nperpcap)，美国受伤人数(nwoundu)，凶手数量（nperps），凶手受伤人数（nwoundte），死亡总数（nkill），受伤总数（nwound）。该数据集的样本数114183，其缺失值的分布如下:

nperps 13166  
nkillus 368  
nkill 4280  
nperpcap 3368  
nwound 8232  
nkillter 2676  
nwoundte 4505  
nwoundus 577

我们对缺失值采用均值的填充方法。采用聚类方法，得出5个随机聚类块。

之后，对属性设置权重，在此，采用熵法确定权重，发生概率小的事件,不肯定程度就越大,发生概率大的事件,不肯定程度就越小,小概率事件发生了,信息量就越大。

设事件 X =, 则有熵函数,式中n是一个正的常数, pi 是事件xi发生的概率。当时， 达到最大值,即当 pi 的值越“集中”,则 的值越大。熵函数与权重成反比,定义 出现的概率为

则属性j的熵 Ej 的计算式为

就某一属性来说,如果在一属性上的表现相当接近,则在这一属性上所占的比率 pij 越集中,此时熵 Ej 较大; 当在这一属性上的值“相同”时,该属性的作用便不太突出,即所对应 的权重ω就越小;反之,若属性指标间的差异越大,即 pij 越分散,熵 Ej 越小时,该属性也就显得越重要,对应的ω值越大[2]。

已知权重须满足及

已知权重ωj与熵 Ej 的关系是成反比的,故令

若决策者主观上对属性j已经有一个先验的权重wjo，则可以将两者结合起来形成一个新的权重

依据上述属性的权重确定值，计算随机聚成5类的概率，概率大者危害等级较高。

把事件按照危害程度分成五级，并且这二十年来危害程度最高的十大恐怖事件为：

|  |
| --- |
| 1 200109110004 |
| 2 200409010002 |
| 3 200210230004 |
| 4 200403110001 |
| 5 200507070001 |
| 6 201107220011 |
| 7 200811260003 |
| 8 201710010018 |
| 9 199808070002 |
| 10 200210120004 |

表1

|  |  |
| --- | --- |
| 事件编号 | 危害级别 |
| 200108110012 | 1 |
| 200511180002 | 2 |
| 200901170021 | 1 |
| 201402110015 | 5 |
| 201405010071 | 4 |
| 201411070002 | 5 |
| 201412160041 | 1 |
| 201508010015 | 5 |
| 201705080012 | 2 |

表2

问题2的数学模型

对2015、2016犯罪集团（每一个集团的攻击次数）数据集分为测试集和验证集，测试集按照70%和30%比例分为测试集和验证集，之后采用分类精度较高的随机森林分类模型，用17个分类变量和7个计算特征作为判断属性。

随机森林(RF)是一种统计学习理论,它是利用bootsrap重抽样方法从原始样本中抽取多个样本,对每个bootsrap样本进行决策树建模,然后组合多棵决策树的预测,通过投票得出最终预测结果根据下列算法而建造每棵树。

其中，用N来表示训练用例（样本）的个数，M表示特征数目。输入特征数目m，用于确定决策树上一个节点的决策结果（ m应远小于M），从N个训练用例（样本）中以有放回抽样的方式，取样N次，形成一个训练集（即bootstrap取样），并用未抽到的用例（样本）作预测，评估其误差。

对于每一个节点，随机选择m个特征，决策树上每个节点的决定都是基于这些特征确定的。根据这m个特征，计算其最佳的分裂方式。

随机森林基于每一个决策树来实行的，在决策树中，随着划分过程的不断进行，希望决策树中的分支节点所包含的样本数尽可能属于同一类别，即节点纯度越来越高，用信息熵度量样本集合纯度,假定当前样本集合D中第k类样本所古的比例为,则D的信息熵定义为：

Ent(D)的值越小，则D的纯度越高。

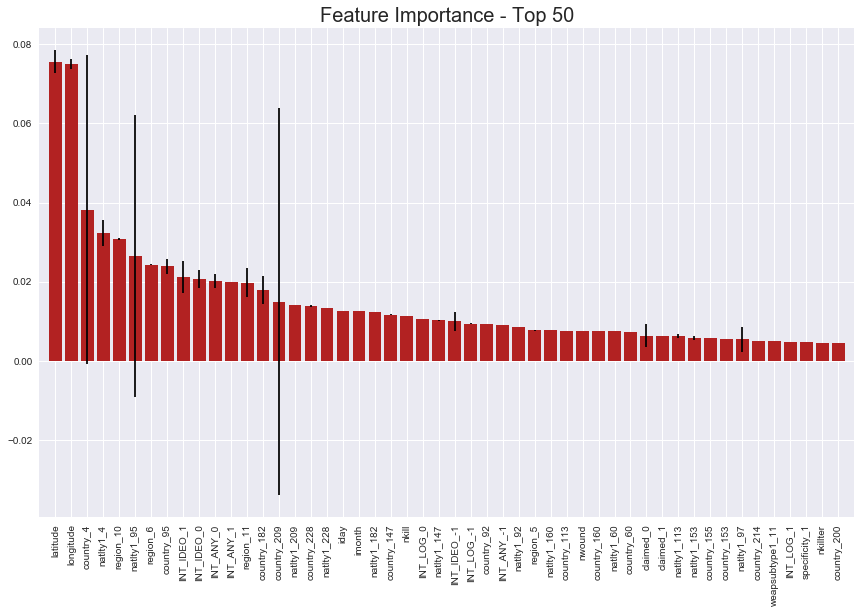
假定离散属性a有V个可能的取值{al, a2 ,…,aV },若使用a来对样本集 D进行划分，则会产生v个分支结点，其中第0个分支结点包含了D中所有在 属性a上取值为aV的样本，记为DV.再考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重。

计算出用属性a对样本集D进行划分所获得的“信息增益率”(information gain)，这里减少信息准则对可取数值数目较多的属性偏好，因此采用信息增益率。

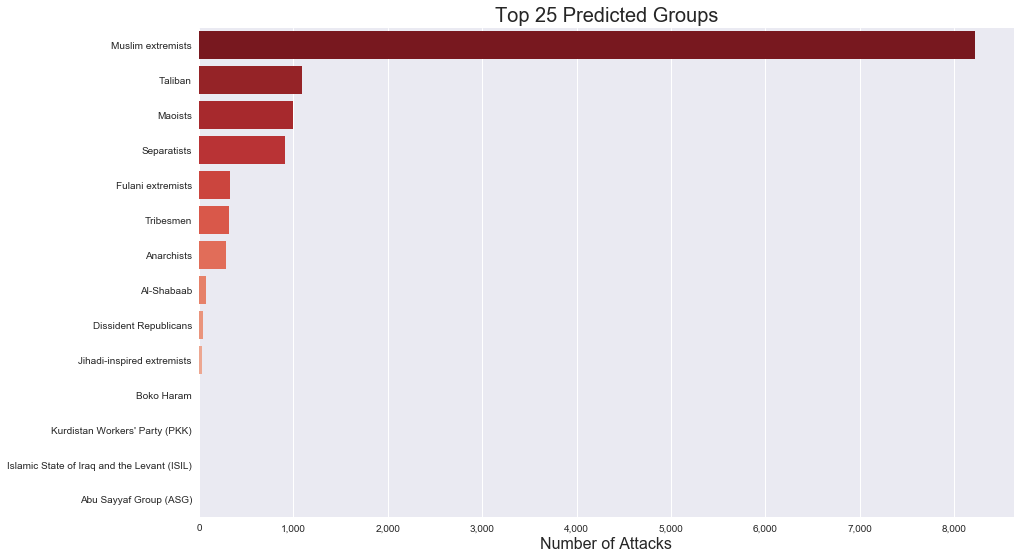
其中

通常属性a的可能性数目越多（V越大），则的值就会越大[3]。

本题目中，显然，对每一个属性计算相应的信息增益，给出基于各个属性根节点进行划分。



图一



图二

对数据处理，选取重要的前50个特征，然后进行预测，可得到排名前25的组织，其中排名前5 的组织为： 1 Muslim extremists : 8218  
 2 Taliban : 1087  
 3 Maoists : 1001  
 4 Separatists : 916  
 5 Fulani extremists : 328

采用随机森林分类算法，对案件进行分类预测，分类结果达94%，说明模型合理性，进而利用Softmax回归模型，对每个样例输出一个概率值，选出前5个对应的概率值，判断犯罪嫌疑人的可能性。

**表2 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 | 1 |  |  |  |  |
| 201702210037 |  | 1 |  |  |  |
| 201703120023 | 1 |  |  |  |  |
| 201705050009 |  |  |  |  | 1 |
| 201705050010 |  |  |  |  | 1 |
| 201707010028 |  |  |  |  | 1 |
| 201707020006 | 1 |  |  |  |  |
| 201708110018 |  | 1 |  |  |  |
| 201711010006 |  | 1 |  |  |  |
| 201712010003 | 1 |  |  |  |  |

表3

采用随机森林模型，可以处理大量的输入变量，决定类别时，评估变量的重要性，可以在内部对于一般化后的误差产生不偏差的估计，更重要的是可以估计遗失的资料，当有很大一部分的资料遗失，仍可以维持准确度，对于不平衡的分类资料集来说，它可以平衡误差。

因此，如果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串联起来统一组织侦査，有助于提高破案效率，有利于尽早发现新生或者隐藏的恐怖分子。

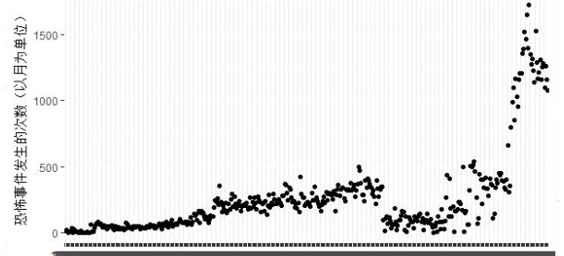
问题3的数学模型

数据预处理

日期数据处理

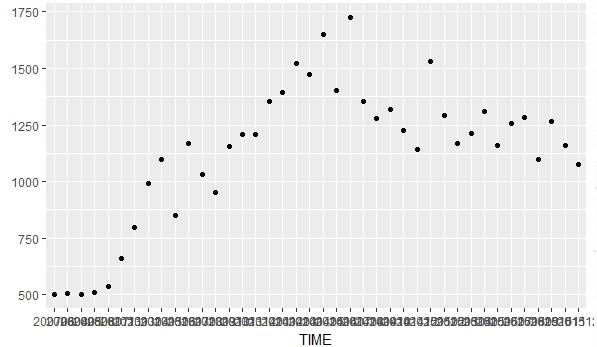
将分开的年月日数据进行合并处理，

反恐多年，呈现的发展态势：



发展态势

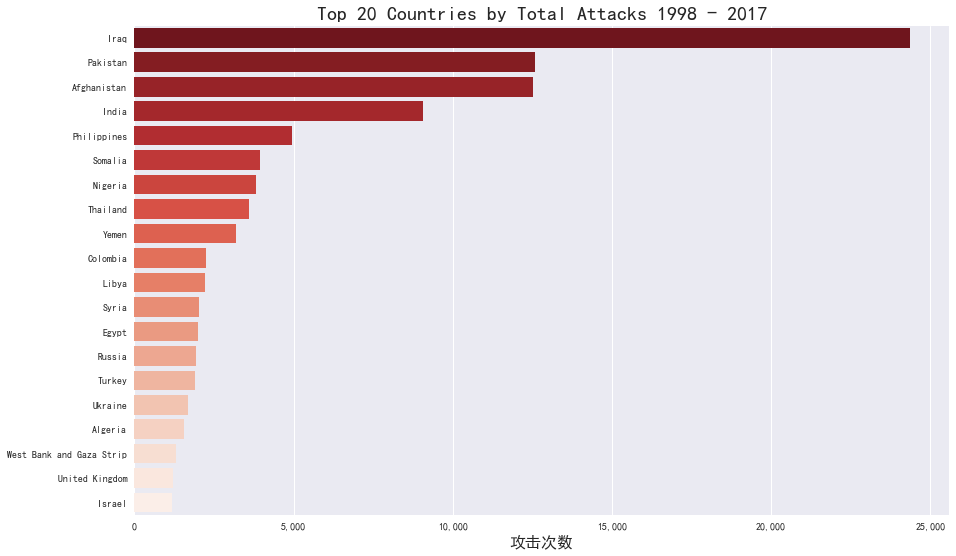
图三



图四

从图三可以看出恐怖袭击事件在前期是以缓慢增长的态势，在中期有个极具下降的趋势，然而好景不长，在之后，能看出恐怖袭击案件是以指数增长趋势发展。

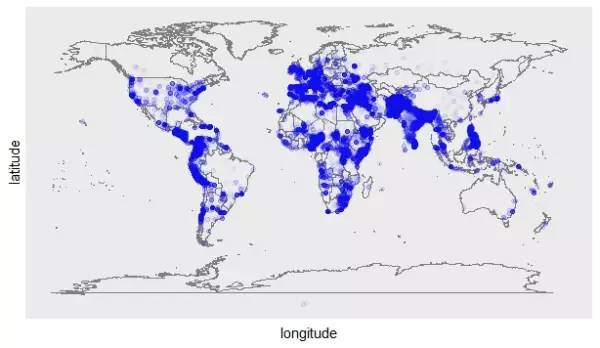
我们可以局部放大下，图四是从2007到2015案件发展散点图。后期呈现下降趋势，反恐计划还是有所收获的。



图五

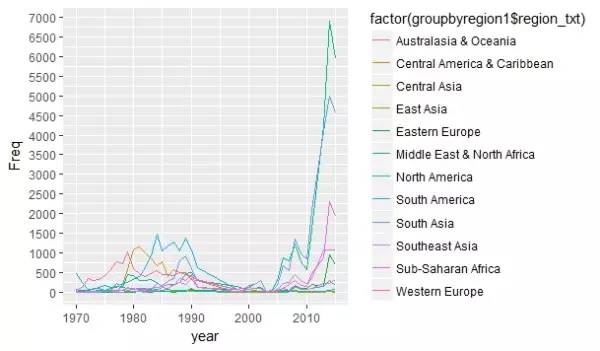
图五可以得到全球恐怖主义指数排行，从1998-2017年受恐怖主义影响排名前20被攻击的国家，Iraq, Pakistan ,Afghanistan, India, Phillippines等。与查阅资料显示结果基本一致，恐怖袭击发生在这些发达国家的部分原因是，这些国家政局动荡，世界上最致命的两个恐怖组织，“博科哈拉姆”（博科圣地）和ISIS，一直在试图扩大在其他国家和地区的活动，这激励我们要加强自身综合国力，加强防范，增强地区反恐合作。在国际和地区层面上协调立场，共同行动，推动建立周边地区的反恐机制。

全球恐怖袭击事件主要分布区域：



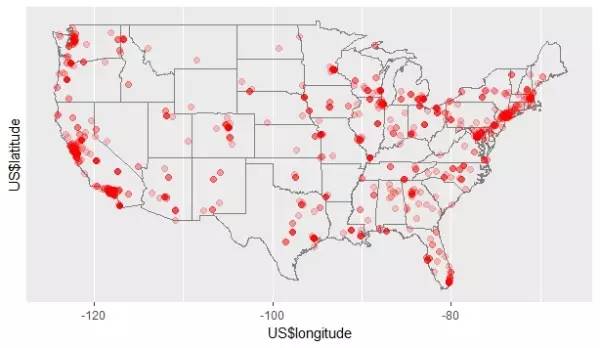
图六

可以看出，恐怖袭击案件最高发地区为东欧，中东南非，西欧，亚洲南。在这些区域当中， 从下图可以看出从2010开始，各地恐怖袭击事件发生的次数明显比之前多。东欧，中东南非，西欧，亚洲东南部恐怖事件发生率比较陡峭。拉高了全球恐怖袭击总体趋势。



图七

从之前的全球分布图，可以看出我国的恐怖袭击事件还是相对较少的。而美国作为世界霸主，它的发生发布情况如下：



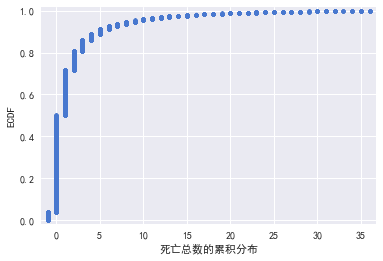
图八

从数据分析结果可以看出，美国东部和加州为恐怖事件高发地区。

全球化是不变的趋势，我国现在也在大力推广“一带一路”政策，所以，在反恐以及防恐的问题上必须明确态度，用一个大国姿态参与全球反恐治理，发挥更大作用，加大对相关国家的反恐能力培训援助，不断提升综合国力。

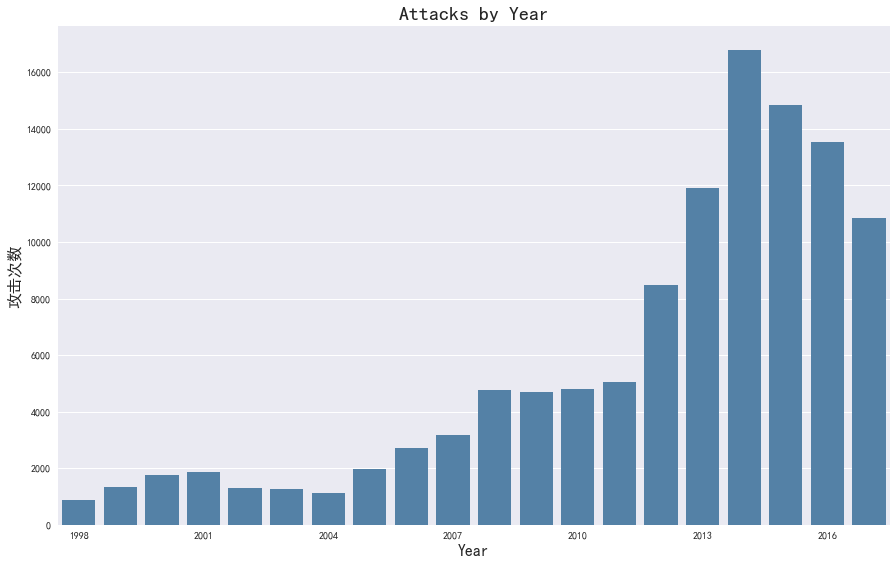
问题4的数学模型

利用附件1的数据研究恐怖袭击主要攻击的手段类型，主要攻击目标和使用的武器，死亡人数和累计分布率，以便做好相应的安保措施和预防手段。

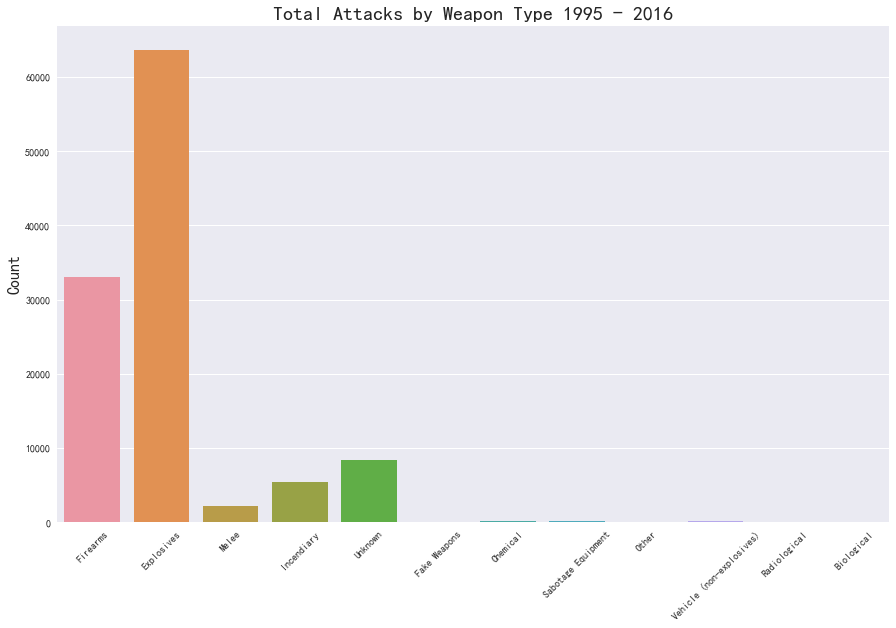


图九

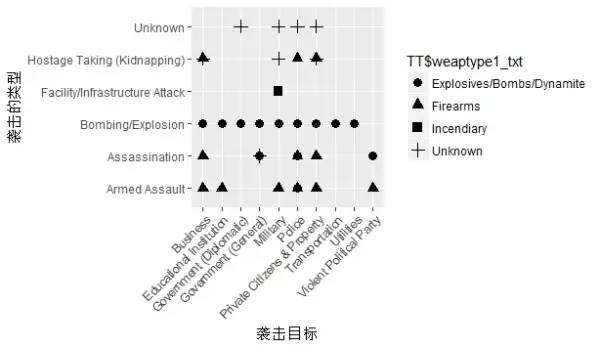
从图九可以看出：死亡总数也在曾上升趋势。



图十



图十一



图十二

从图十至图十二，我们得到：

1）轰炸，爆炸 是犯罪分子最常使用的犯罪武器

2）经常攻击的目标是商业，军队，警方和普通公民。

3）随着年份的变化，攻击次数到2016年之前一直处于增长状态，2.014，2015，2016三年最高，通过采取措施后有所下降。

六、模型评价与推广

反对恐怖主义是我们每个国家应该承担的责任，是我们每个公民应该有的意识。当前及今后相当长一段时期，国际恐怖主义发展新态势也会给中国的反恐带来严重压力，会对我国“一带一路”的持续推进构成重要安全威胁。

本文对恐怖事件记录数据的量化分析，包括事件的危害程度等级，潜在危害因素和组织，以及未来反恐态势的趋势等，发现在各国的共同努力下，防恐还是有所收获，恐怖事件发生次数呈下降趋势，但仍不可掉以轻心，给出具体如下建议：未来一段时期，中国在国际反恐方面应着力加强三个方面的工作:一是切实不断提升自身反恐能力，突破反恐情报搜集与境外反恐行动的瓶颈，实现境内与境外反恐能力的同步发展、反恐行动与情报搜集能力的同步发展;二是扩大国际反恐合作，尤其要有针对性地加大与大国、周边国家以及地区重点国家的反恐合作，同时也要以大国姿态参与全球反恐治理，发挥更大作用，加大对相关国家的反恐能力培训援助。三是针对在我国周边扩张新态势及对我国外溢安全威胁增大，中方应及早谋划，加强防范，增强地区反恐合作。在国际和地区层面上协调立场，共同行动，推动建立周边地区的反恐机制。四：对于网络恐怖主义问题应采取的严厉的打击措施，对恐怖组织或恐怖分子利用互联网发布音频、视频以及实施恐怖行为， 包括煽动、招募、资助或策划等活动[5]。

通过本文的研究也可以应用到城市遭受恐怖袭击的评估研究，保险公司评估和管理恐怖风险暴露以及边疆动乱地带，随着数据的不断更新和完善，模型和评价体系将会更完善。希望可以居安思危通过预防措施，防范于未然，建立一个更好的秩序和更美好的中国。

七、参考文献

[1] 宋海军，浅谈反恐的一些措施，赤子，2013年5月下总第282期供稿

[2] 江南，白小双，孙娟娟，基于多属性决策的统计数据分级评价模型，测绘学报，第36卷第2期，2007.5

[3] 周志华，机器学习决策树，清华大学出版社，73-79

[4] <http://www.sohu.com/a/139172670_716346>

[5] 王娟，网络恐怖主义活动的危害及控制，

<http://www.docin.com/p-740920119.html>