“华为杯”第十五届中国研究生数学建模竞赛

学　　校	北京理工大学
参赛队号	**18100070052**
队员姓名	**1.刘泾洋**
	2.樊振辉
	3.毕晋攀

“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：

大数据时代涌现出的诸多数据处理与挖掘算法，使得对海量廉价数据进行分析并对未来变化趋势进行准确预测成为可能。本文基于 GTD 全球恐怖主义数据库的恐袭事件数据，主要完成了对恐袭事件的分析处理、规律统计与归纳、数据分类及对犯罪组织的预测。首先，通过统计分析数据设计了基于 region 及其他一般特征的两级恐袭事件分级方法，用于提高危害程度分级的准确性与客观性。其次运用 CART 分类树对样本数据进行分类并预测，在此基础上设计了合理的规则实现对不同组织危害程度的划分。随后基于历史统计数据，利用多元统计分析中的因子分析方法，建立了恐怖袭击的风险评估模型。最后，利用相关性分析方法对中国地区的恐怖主义事件进行特征分析，提出了相应的防控措施。

任务一需要利用样本数据为恐袭事件设计客观准确的危害程度分级方法。首先根据 region（地区）特征对 1998-2017 年间的样本数据进行聚类分析，将恐袭事件根据 region 特征划分为 5 个互不交叉的类。在此基础上对每个类选取评判指标进行因子分析，建立了主成分回归模型。最后利用该分级方法统计出近二十年间危害程度最高的 10 次恐袭事件，并且划分了典型事件的危害等级如表 2.20 所示。

任务二需要利用恐袭样本数据对 15-16 年间未确定组织的事件进行分类，并根据危害性大小对分类结果进行排序，筛选出危害程度前 5 的组织来对 17 年的典型事件进行嫌疑程度排序。首先统计并合理推断出组织犯罪的最大时间跨度规律、以及新兴组织的最小犯罪时间间隔规律，利用这两个规律对 1998-2016 年间的数据进行筛选处理，剔除了具有干扰性的样本数据。其次，通过分析每一个特征与不同组织的关系，合理地选取了分类特征。继而，以犯罪组织作为训练标签，采用 CART 分类树对筛选出的样本数据进行训练，从而得到预测模型，5-折交叉验证的准确率达到 86%，表明了该模型具有较高的精度。随后利用该模型对 2015-2016 未知组织的数据进行预测，得到预测结果如附件一。最后，设计了针

对犯罪组织的危害性排序规则并得到危害性最大的 5 组组织如表 3.4 所示，在此基础上设计了嫌疑度排序规则，得到了 2017 年典型事件的嫌疑程度排序如表 3.7。

任务三需要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，进而预测下一年全球或某些重点地区的反恐态势。首先对 2015-2017 年的恐袭事件进行统计分析，选取了最能客观反映恐袭事件风险指标的特征。进而利用多元统计分析中的因子分析法建立了恐怖袭击的风险评估模型，为 2018 年国际反恐态势提供参考。

任务四需要对样本数据进行进一步利用。首先通过相关性分析法分析了中国恐怖主义事件与国际恐怖主义事件的相关性，从主要地区、主要恐怖组织、攻击方式、武器类型等方面研究了中国恐怖主义的时间特征，并在此基础上提出了相应的防控措施。

关键词：因子分析、主成分回归模型、CART 分类树、风险评估模型

目录

第一章 问题重述

1.1 问题背景

恐怖袭击是指极端分子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道义的攻击方式，它不仅具有极大的杀伤性与破坏力，能直接造成巨大的人员伤亡和财产损失，而且还给人们带来巨大的心理压力，造成社会一定程度的动荡不安，妨碍正常的工作与生活秩序，进而极大地阻碍经济的发展。恐怖袭击从 20 世纪九十年代以来，有在全球范围内迅速蔓延的严峻趋势。美国 911 事件是人类历史上最为惨烈的恐怖事件之一，它引起了世界对于恐怖组织的正视，自从 911 事件发生以来，各国已经开始加大力度在反恐的事业中。美国通过了反恐《2001 法案》[1]，中国也将“恐怖主义，极端主义，分裂主义”三股势力列入要坚决打击的范围，但是恐怖主义仍然广泛存在，并且随着时间的增长正在不断地发展壮大，时刻威胁着人们的日常生活，如何反恐成为一个被世界各界广泛关注的问题。

恐怖主义是人类的共同威胁，打击恐怖主义是每个国家应该承担的责任。对恐怖袭击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，为反恐防恐提供有价值的信息支持。

1.2 需要研究的问题

问题一：对灾难性事件比如地震、交通事故、气象灾害等等进行分级是社会管理中的重要工作。通常的分级一般采用主观方法，由权威组织或部门选择若干个主要指标，强制规定分级标准，对灾难性事件比如地震、交通事故、气象灾害等等进行分级是社会管理中的重要工作。通常的分级一般采用主观方法，由权威组织或部门选择若干个主要指标，强制规定分级标准。

为了提高量化分级的合理性和准确度，我们需要对样本数据进行分析，结合样本特征找出一种能够真实反映恐怖袭击事件危害性等级的划分标准，进而完成对典型事件的危害性等级划分。

问题二：样本数据中有多起恐怖袭击事件尚未确定作案者。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查，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。我们需要对已明确的恐怖袭击事件进行分类，利用分类模型去预测未明确恐怖袭击事件的发动者信息。此外还需要根据样本特征设计合理的恐怖组织危害程度分级方法以及合理的嫌疑度排序方法，从而可以实现对某一具体恐怖袭击事件的组织嫌疑度排名。

问题三：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。通过对样本数据进行分析，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。

问题四：需要对问题进行进一步地利用。通过对数据进行分析，研究中国恐怖主义事件的特征，进而研究中国恐怖注意事件随年变化与国际恐怖主义事件发展变化之间的相关性，并且应该针对这些特征给出对应的防控措施。

第二章 任务一解答

在本章中，我们首先选取了影响恐怖袭击事件的危害程度的评判指标，根据恐怖袭击地点(region)进行了聚类分析，将 1998 年~2017 年的恐怖袭击事件分成了五类。随后对每一个类别将选取的评判指标进行因子分析，建立了相对应的主成分回归模型，利用主成分回归模型对近二十年的恐怖袭击事件按照危害程度进行排名，得到了十大危害的恐怖袭击案事件。最终对表格中的事件完成了分类[1]。

2.1 模型的符号说明

以下符号所代表的评判恐怖袭击事件危害程度的指标，是本文所有问题研究的基础指标。具体形式如表 2.1 所示。

表 2.1 模型符号

符号	含义
x_1	入选标准 1 (crit1)
x_2	入选标准 2 (crit2)
x_3	入选标准 3 (crit3)
x_4	疑似恐怖主义 (doubtterr)
x_5	成功的攻击 (success)
x_6	攻击类型 (attacktype1)
x_7	凶手数量 (nperps)
x_8	武器类型 (weapontype1)
x_9	死亡总数 (nkill)
x_{10}	受伤总数 (nwound)
x_{11}	财产损害程度 (propextent)

由于 GTD 提供的恐怖袭击事件数据中，有些项是文本格式。因此，需要对文本资料进行分析整理为可评价数据。我们选取的防空数据包含特征属性 11 项，分别为入选标准 1~3、疑似恐怖主义、成功的攻击、攻击类型、凶手数量、武器类型、死亡总数、受伤总数以及财产损失程度。下面我们对上述选取的指标依次指定评价标准：

1、入选标准 1~3

这是分类数据类型，用 1 表示该恐怖袭击事件满足该标准，0 表示该恐怖袭击事件不满足该标准。我们认为当恐怖袭击事件被判定满足这三个标准时，恐怖袭击事件的危害程度比较大。因此，以 0 和 1 来量化此评判指标。

2、疑似恐怖主义

GTD 数据中不能确定一个事件是否符合所有入选的标准，因此本指标就是用

来评判该具体事件属于恐怖袭击事件的可能性。我们采用的量化标准如下：0 代表怀疑该事件是恐怖主义行为，1 代表基本上不怀疑该事件是恐怖主义行为，即确定该事件属于恐怖主义。

3、成功的攻击

该指标用来衡量恐怖袭击事件是否攻击成功，成功的危害程度是大于失败的危害程度的。因此量化标准如下：1 表示袭击成功，0 表示袭击失败。

4、攻击类型

该指标为文本数据类型，攻击类型的层次共分为 9 种，包括：暗杀、劫持、绑架、路障事件、轰炸/爆炸、未知、武装突袭、徒手攻击、设施/基础设施攻击。量化标准如下：我们依次将上述攻击类型按照危害程度分为 9 个等级，1 级表示危害程度最轻，9 级表示危害程度最大。即暗杀为 9 级，设施/基础设施攻击为 1 级。

5、凶手数量

该评价指标为数值数据类型，凶手数量越多，会造成越危险的恐怖袭击，因此凶手数量指标和恐怖袭击事件成正比。在 GTD 数据库中，详细记录了参与这一事件的恐怖分子总数，当凶手的数量没有被报道是，会出现“-99”或“未知”。我们认为恐怖袭击事件危害性越大，越会受到媒体的关注，政府反恐的力度也会很大，因此恐怖分子的数量越会被报道出来。因此当凶手数量为“-99”时，可以看出该恐怖事件的影响力较小，所以量化标准如下：当凶手数量为“-99”时，我们将其设为 0，其余的情况就是真实的数量。

6、武器类型

该评价指标为分类数据类型，武器类型共分为 13 种，包括：生物武器、化学武器、放射性武器、核武器、轻武器、爆炸物/炸弹/炸药、假武器、燃烧武器、致乱武器、交通工具、其它、未知。量化标准如下：我们依次将上述武器类型按照危害程度分为 13 个等级，1 级表示危害程度最轻，13 级危害程度最大，即用 1 表示“未知”武器，用 13 表示生物武器。

7、死亡总数

该指标为数值数据类型，存储了事件中所有确认死亡的总人数，当数字模糊不清时，这个指标记录为空值。量化标准如下：当凶手数量为空值时，我们将其设为 0，其余的情况就是真实的数量。

8、受伤总数

该指标为数值数据类型，存储了事件中已证实的受到非致命伤害的数量，当数字模糊不清时，这个指标记录为空值。量化标准如下：当凶手数量为空值时，我们将其设为 0，其余的情况就是真实的数量。

9、财产损失程度

该指标为数值数据类型，共分为 4 种类型：灾难性的、重大的、较小的和未知。我们采用的量化标准如下：依次分为四个等级，1 表示危害程度最轻，4 表示危害程度最重，即 1 表示未知，4 表示灾难性的。

2.2 对恐怖袭击事件进行聚类分析

我们在本节对附件 1 中的恐怖袭击事件按照危害程度进行了聚类分析，旨在能够确定不同的恐怖袭击事件的严重等级，对后面的分析过程奠定了基础。

2.2.1 聚类分析

我们针对不同的 12 个地域进行了划分，分别为 Australasia & Oceania(AO)、East Asia(EA)、Central America & Caribbean(CAC)、Central Asia(CA)、East Europe(EE)、North America(NA)、Middle East & North Africa(MENA)、South America(SA)、South Asia(SAS)、Sub-Saharan Africa(SSA)、Southeast Asia(SEA)、Western Europe(WE)。我们对附件 1 中的数据进行预处理，在不断尝试之后，选择了以下指标作为体现恐怖袭击危害程度的指标：

每个地区发生的恐怖袭击总数、每个地区成功的恐怖袭击总数、每个死亡总数、每个地区的受伤总数、每个地区的财产损失程度、每个地区使用的武器类型、每个地区的凶手总数、每个地区的袭击类型、每个地区的入选标准 1~3 以及每个地区的疑似恐怖袭击标准，分别定义为变量 $sum\_x_1 \sim sum\_x_{12}$。

在确定指标之后，我们对表 2.1 中的量化数据分别按地区进行求和，得到变量 $sum\_x_1 \sim sum\_x_{12}$，所得数据如表 2.2 所示。

表 2.2 聚类数据

	sum_x_1	sum_x_2	sum_x_3	sum_x_4	sum_x_5	sum_x_6
AO	85	21	51	44	518	79
CAC	98	106	173	31	706	109
CA	277	299	675	100	1890	404
EA	211	914	1488	86	1327	3529
EE	4256	6435	10965	1320	28948	10625
MENA	41640	119126	188887	11290	276481	32304
NA	741	3530	18665	421	4557	1192
SA	2750	4051	5271	906	17309	14421
SAS	37573	76859	110495	11434	248270	130865
SEA	9701	9604	20053	3254	72484	50407
SSA	13274	58768	40416	3950	84253	53854
WE	3577	1062	6396	1801	22272	4180

	sum_x_7	sum_x_8	sum_x_9	sum_x_{10}	sum_x_{11}	sum_x_{12}
AO	68	307	78	85	84	14
CAC	80	629	96	97	98	11
CA	237	1589	267	277	274	35

EA	177	1131	203	211	208	34
EE	3669	24856	4155	4252	3385	1271
MENA	36533	241046	41469	41438	35777	7092
NA	611	3098	673	740	737	154
SA	2489	14618	2722	2747	2588	318
SAS	32413	211974	37205	37390	34152	4689
SEA	9172	60799	10147	10395	9018	2013
SSA	12234	72083	13136	13201	11110	2680
WE	2731	17527	3516	3564	3559	278

通常聚类分析算法一般包含四个部分:(1)特征获取与选择;(2)计算相似度;(3)分组;(4)聚类结果展示。

本文所采用聚类算法为 K-Means 均值聚类分析方法。常见的 K-Means 算法的工作流程是这样的：首先，随机确定 k 个初始点作为簇的质心。然后将数据集中的每个点分配到一个簇中，具体来讲，就是为每个点找距离其最近的质心，并将其分配给该质心所对应的簇。这一步完成之后，将每个簇的质心更新为该簇所有点的平均值。这个过程不断重复，直到准则函数收敛。通常，采用平方误差准则[1]。

$$E = \sum_{k} \sum |p - m_i|^2$$

式中，E 数据集中所有对象的平方误差和，p 是空间中的点，表示给定的对象，m_i 是簇 c_i 的均值，也即是质心。

K-Means 算法的整个过程如下：划分的 K-Means 算法是基于簇中对象的平均值。输入：簇的数目 k 以及包含所有对象的数据集；输出：k 个簇，其中平方误差准则最小方法。具体流程如下[2]：

1) 随机选择 k 个对象作为初始的簇中心；
2) Repeat；
3) 将对象赋给与其距离最近的簇；
4) 根据每个簇中所有对象的平均值，更新簇的中心；
5) 直到簇中心不再发生变化。

2.2.2 聚类分析结果及说明

由于原始数据过大，我们才对数据进行了预处理得到了表 2.1 所见的聚类数据的形式。对于上述数据进行聚类分析可以得到对应的聚类结果。

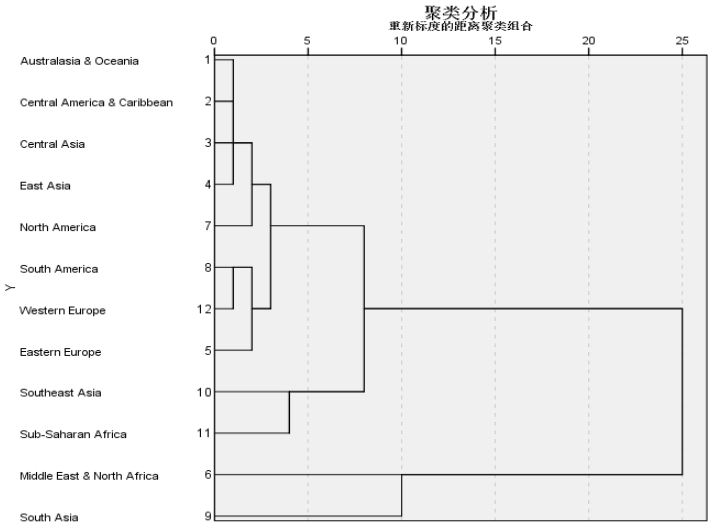

图 2.1 恐怖袭击事件聚类图

表 2.3 最终聚类中心之间的距离

聚类数	1	2	3	4	5
1	0	425479.771	97372.258	131175.184	370446.070
2	42549.771	0	345617.333	308752.633	139153.283
3	97372.258	345617.333	0	56158.047	276432.140
4	131175.184	308752.633	56158.047	0	245750.180
5	370446.070	139153.283	276432.140	245750.180	0

图 2.1 为恐怖袭击事件的聚类图，表 2.3 为最终聚类中心之间的距离。从上图可以看出，最终附件 1 中的恐怖袭击事件最终被分为了 5 类，且不同类之间的距离较远，聚类效果较好。分类结果如表 2.4 所示。

表 2.4 分类结果

类别	地区
一类	AO、CAC、CA、EA、EE、NA、SA、WE
二类	MENA
三类	SAS
四类	SEA
五类	SSA

2.3 因子分析

因子分析是简化、分析高维数据的一种统计方法。可在许多变量中找出隐藏的具有代表性的因子，将相同本质的变量归入一个因子，可减少变量的数目，还可检验变量间关系的假设[3]。

假设 p 维随机变量 $X=\left(X_1,X_2,\ldots,X_p\right)^T$ 满足：

$$X=\mu+A\bar{f}=\bar{e}$$

其中 $\bar{f}=\left(f_1,f_2,...,f_q\right)^T$ 是 q 维随机变量，$q\le p$，满足 $E\bar{f}=0,E\bar{f}\bar{f}^T=\bar{I}_q$，它的分量 f_i 称为公共因子，对 X 的每个分量都起作用。$\bar{e}=\left(e_1,e_2,...,e_p\right)^T$ 是 p 维不可观测的随机向量，满足

$$E\bar{e}=0,E\bar{e}\bar{e}^T=diag\left(\sigma_1^2,\sigma_2^2,...,\sigma_p^2\right)$$

且 $E\bar{f}\bar{e}^T=0$，e 的分量 $\bar{e_i}$ 成为特殊因子。

μ 和 A 为参数。若 X 满足上式，则称随机向量 X 具有因子结构。其中

$$A=\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1q} \\ a_{21} & a_{22} & \ldots & a_{2q} \\ \ldots & \ldots & & \ldots \\ a_{p1} & a_{p2} & \ldots & a_{pq} \end{bmatrix}$$

由模型及其假设前提可知，公共因子 $f_1,f_2,...,f_q$ 相互独立且不可测，是在原始变量的表达式中出现的因子。各特殊因子之间以及特殊因子与所有的公共因子之间也都是相互独立的。矩阵 A 中的元素 a_{ij} 称为因子载荷，a_{ij} 的绝对值越大（$|a_{ij}|\le 1$），表明 X_i 和 f_j 的相依程度越大，或称为公共因子 f_j 对 X_i 的载荷量越大。

我们对数据进行了整体的因子分析，可以得到各个变量的公因子方差，公因子方差主要就是看几个公因子方差的累计贡献率，累计贡献率越高，说明提取的这几个公因子对于原始变量的代表性或者说解释率越高，整体的效果就越好。累计贡献率越低，说明提取的公因子的代表性或者说解释率越差，整体的效果就越差。我们认为其实当公因子方差大于 60%-70%的时候，我们就可以接受了。

下面我们对上述聚类分析所获得的 5 个类分别进行因子分析。

2.3.1 第一类

表 2.5 第一类各主成分的方差贡献表

成分	特征值	方差百分比(%)	累计贡献率(%)
1	1.969	17.901	17.901
2	1.856	16.877	34.778
3	1.610	14.632	49.411
4	1.177	10.696	60.107

5	1.035	9.408	69.515
6	1.000	9.089	78.603
7	0.997	9.065	87.668
8	0.674	6.124	93.792
9	0.394	3.580	97.372
10	0.227	2.065	99.438
11	0.062	0.562	1

表 2.4 所示为第一类因子分析的方差贡献表，从中可以看出前五个成分的累积方差贡献率达到了接近 70%，即前五个成分所包含的信息占据了全部信息的 70%，可以用它们来代表整体变量，我们选取前五个成分作为主成分。各变量在前五个主成分中的系数见表 2.6。

表 2.6 各变量在主成分中的系数 (第一类)

	x_1	x_2	x_3	x_4	x_5	x_6
1	0.103	0.026	0.27	-0.29	0.134	-0.23
2	-0.26	-0.07	-0.67	0.724	-0.08	-0.08
3	0.21	0.046	0.43	-0.50	-0.37	-0.37
4	-0.33	-0.17	0.122	0.104	0.656	0.656
5	0.845	-0.09	-0.39	-0.05	0.336	0.336

	x_7	x_8	x_9	x_{10}	x_{11}
1	0.035	-0.212	0.887	0.891	0.318
2	-0.02	0.357	0.415	0.405	-0.370
3	-0.06	0.704	0.068	0.07	-0.152
4	-0.06	0.360	-0.06	-0.066	0.617
5	0.16	0.062	-0.013	-0.026	0.086

上表为 11 个变量在前五个主成分中的系数，可以看出，第一主成分主要由 x_9 和 x_{10} 决定，第二主成分主要由 x_4 决定，第三主成分主要由 x_6 和 x_8 决定，第四主成分主要由 x_5 和 x_{11} 决定，第五主成分主要由 x_1 决定。

2.3.2 第二类

表 2.7 第二类各主成分的方差贡献表

成分	特征值	方差百分比(%)	累计贡献率(%)
1	2.057	18.698	18.698
2	1.683	15.302	34.000
3	1.325	12.048	46.048
4	1.071	9.733	55.781
5	1.026	9.327	65.108
6	0.998	9.071	74.179
7	0.974	8.858	83.038
8	0.854	7.765	90.803
9	0.704	6.401	97.203
10	0.234	2.129	99.333
11	0.073	0.667	1

表 2.7 所示为第二类因子分析的方差贡献表，从中可以看出前五个成分的累积方差贡献率达到了 65%，我们选取前五个成分作为主成分。各变量在前五个主成分中的系数见表 2.8。

表 2.8 各变量在主成分中的系数 (第二类)

	x_1	x_2	x_3	x_4	x_5	x_6
1	0.107	0.145	0.822	-0.85	0.107	0.492
2	-0.05	0.015	-0.47	0.466	0.002	0.793
3	0.008	-0.09	-0.08	0.100	0.382	-0.040
4	-0.40	-0.23	0.051	0.122	0.633	-0.080
5	0.060	-0.70	-0.39	-0.004	-0.108	0.051

	x_7	x_8	x_9	x_{10}	x_{11}
1	-0.028	0.513	-0.045	0.130	0.313
2	-0.006	0.781	-0.063	-0.015	-0.030
3	0.045	0.016	0.731	0.754	0.218
4	-0.025	-0.020	-0.321	-0.206	0.539
5	0.701	0.049	0.016	-0.043	-0.093

上表为 11 个变量在前五个主成分中的系数，可以看出，第一主成分主要由 x_3 和 x_4 决定，第二主成分主要由 x_6 和 x_8 决定，第三主成分主要由 x_9 和 x_{10} 决定，第四主成分主要由 x_5 决定，第五主成分主要由 x_2 和 x_7 决定。

2.3.3 第三类

表 2.9 第三类各主成分的方差贡献表

成分	特征值	方差百分比(%)	累计贡献率(%)
1	1.934	17.584	17.584
2	1.710	15.543	33.127
3	1.556	14.146	47.273
4	1.158	10.528	57.801
5	1.015	9.225	67.026
6	1.006	9.146	76.172
7	0.989	8.992	85.164
8	0.789	7.171	92.335
9	0.471	4.281	96.616
10	0.284	2.584	99.200
11	0.088	0.800	1

表 2.10 所示为第三类因子分析的方差贡献表，从中可以看出前六个成分的累积方差贡献率达到了 76%，，我们选取前六个成分作为主成分。各变量在前六个主成分中的系数见表 2.10。

表 2.10 各变量在主成分中的系数 (第三类)

	x_1	x_2	x_3	x_4	x_5	x_6
1	0.233	0.177	0.887	-0.95	-0.045	0.206
2	-0.08	-0.02	-0.17	0.194	-0.023	0.881
3	0.071	0.037	0.065	-0.08	0.307	-0.192
4	-0.27	-0.10	0.007	0.109	0.711	-0.013
5	0.604	0.581	-0.34	-0.013	0.107	0.009
6	0.630	-0.77	-0.03	-0.002	0.096	0.026

	x_7	x_8	x_9	x_{10}	x_{11}
1	-0.039	0.228	-0.118	-0.017	0.236
2	0.004	0.878	0.210	0.214	-0.012
3	0.190	-0.152	0.811	0.807	0.198
4	-0.021	0.085	-0.209	-0.174	0.691
5	0.369	0.048	-0.072	-0.137	0.159
6	0.089	0.030	-0.019	-0.049	0.016

上表为 11 个变量在前六个主成分中的系数，可以看出，第一主成分主要由 x_3 和 x_4 决定，第二主成分主要由 x_6 和 x_8 决定，第三主成分主要由 x_9 和 x_{10} 决定，第四主成分主要由 x_5 和 x_{11} 决定，第五主成分主要由 x_1 决定，第六主成分主要由 x_2 决定。

2.3.4 第四类

表 2.11 第四类各主成分的方差贡献表

成分	特征值	方差百分比(%)	累计贡献率(%)
1	1.970	17.907	17.907
2	1.686	15.326	33.233
3	1.424	12.949	46.182
4	1.114	10.130	56.132
5	1.004	9.126	65.438
6	1.000	9.087	74.526
7	0.979	8.902	83.428
8	0.813	7.389	90.817
9	0.614	5.580	96.397
10	0.303	2.759	99.155
11	0.093	0.845	1

表 2.11 所示为第四类因子分析的方差贡献表，从中可以看出前五个成分的累积方差贡献率达到了 65%，我们选取前五个成分作为主成分。各变量在前五个主成分中的系数见表 2.12。

表 2.12 各变量在主成分中的系数 (第四类)

	x_1	x_2	x_3	x_4	x_5	x_6
1	-0.26	-0.07	-0.763	0.832	-0.121	0.509
2	0.110	0.044	0.468	-0.487	-0.055	0.750
3	-0.05	-0.003	-0.061	0.085	0.390	-0.045
4	-0.71	-0.133	0.218	0.123	0.442	0.073
5	-0.22	0.975	-0.019	-0.011	-0.015	-0.013

	x_7	x_8	x_9	x_{10}	x_{11}
1	-0.052	0.436	0.092	0.010	-0.382
2	-0.099	0.762	0.153	0.133	-0.139
3	0.122	-0.047	0.758	0.763	0.288
4	-0.001	0.177	-0.207	-0.214	0.452
5	0.046	-0.008	-0.003	0.005	-0.032

上表为 11 个变量在前五个主成分中的系数，可以看出，第一主成分主要由 x_4 决定，第二主成分主要由 x_6 和 x_8 决定，第三主成分主要由 x_9 和 x_{10} 决定，第四主成分主要由 x_1 决定，第五主成分主要由 x_2 决定。

2.3.5 第五类

表 2.13 第五类各主成分的方差贡献表

成分	特征值	方差百分比(%)	累计贡献率(%)
1	2.036	18.511	18.511
2	1.677	15.243	33.754
3	1.295	11.773	45.527
4	1.056	9.600	55.126
5	1.015	9.227	64.353
6	1.005	9.135	73.488
7	0.993	9.025	82.513
8	0.864	7.856	90.369
9	0.725	6.591	96.960

10	0.264	2.399	99.359
11	0.071	0.641	1

表 2.13 所示为第五类因子分析的方差贡献表，从中可以看出前六个成分的累积方差贡献率达到了 74%，，我们选取前六个成分作为主成分。各变量在前六个主成分中的系数见表 2.14。

表 2.14 各变量在主成分中的系数 (第五类)

	x_1	x_2	x_3	x_4	x_5	x_6
1	0.116	0.106	0.848	-0.868	0.056	0.422
2	-0.14	-0.06	-0.39	0.421	-0.041	0.819
3	-0.04	0.09	-0.05	0.081	0.274	-0.104
4	0.546	0.290	-0.08	-0.127	-0.571	0.080
5	0.766	-0.30	-0.14	-0.011	0.430	0.027
6	0.078	0.887	-0.21	-0.008	0.359	-0.013

	x_7	x_8	x_9	x_{10}	x_{11}
1	0.051	0.458	0.064	0.069	0.369
2	0.010	0.805	0.084	0.042	-0.015
3	0.305	-0.074	0.765	0.683	0.203
4	0.210	-0.007	0.161	0.179	-0.466
5	-0.356	0.048	-0.032	0.012	0.045
6	-0.040	0.012	-0.044	-0.034	0.185

上表为 11 个变量在前六个主成分中的系数，可以看出，第一主成分主要由 x_3 和 x_4 决定，第二主成分主要由 x_6 和 x_8 决定，第三主成分主要由 x_9 和 x_{10} 决定，第四主成分主要由 x_5 决定，第五主成分主要由 x_1 决定，第六主成分主要由 x_2 决定

2.4 主成分回归模型的建立

上一节中我们对通过聚类分析获得的五个类的 11 个指标进行了因子分析，获得了我们所需要的主成分，消除了指标间的多重共线性。本节我们针对每个主成分所对应的变量建立了主成分回归模型，成功地建立了评判恐怖袭击事件危害程度的数学模型。

2.4.1 模型假设

(1) 假设国际政治、经济格局不会发生较大改变，国际安全形势较为平稳，局部地区不会发生大规模武装冲突。

(2) 假设不考虑宗教、信仰格局变化以及国际教育文化水平变化。

(3) 假设不考虑国家及地区发生的大规模战争。

2.4.2 模型的建立

本节需要考虑不同的恐怖袭击事件的危害程度，我们建立线性回归模型来进行评价，回归模型的数值越大，代表该事件的危害程度越大。线性回归模型的形式如下所示：

$$y=\sum_{i=1}^{n}\omega_i x_i=\omega_1 x_1+\omega_2 x_2+\cdots+\omega_n x_n$$

式中，ω代表每个变量所对应的权值，权值越大，代表该变量对恐怖袭击事件的危害程度影响越大。x为前面量化过后的评价标准。

权值ω的确定对于模型是否精确有着十分重要的影响，确定数据的权重是进行数据分析的重要前提。一旦权值的设定不符合实际情况，将会对最后的结果造成很大的偏差。本文通过因子分析的结果来确定权重。具体步骤如下：

(1) 对我们选择的指标所对应的数据进行因子分析(主成分方法)，使用方差最大化旋转;

(2) 写出主因子得分和每个主成分的方程贡献率。

$$F_j=\beta_{1j}*x_1+\beta_{2j}*x_2+\cdots+\beta_{nj}*x_n$$

其中F_j为主成分($j=1,2,\ldots,m$)，$x_1,x_2,\ldots x_n$为各个评价指标，$\beta_{1j},\beta_{2j},\beta_{3j},\ldots,\beta_{nj}$为各指标在主成分$F_j$中的系数得分，用$e_j$表示$F_j$的方程贡献率。

(3) 求出指标权重

$$\omega_i=\frac{\sum_{i=1}^{m}\beta_{ij}e_j}{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_{ij}e_j}$$

2.4.3 主成分回归模型的求解

1、第一类

根据 2.3 节所做的因子分析，我们可以得到五个主成分，主成分可由x_1、x_4、x_5、x_6、x_8、x_9、x_{10}和x_{11}决定，因此我们可以得到y关于x_1、x_4、x_5、x_6、x_8、x_9、x_{10}和x_{11}的主成分回归函数。根据前文所提出的权重确定方法，所得到的各变量的权重值如下表所示。

表 2.15 第一类变量所对应的权重值

变量	x_1	x_4	x_5	x_6	x_8	x_9	x_{10}	x_{11}
权重	0.0512	0.003	0.061	0.179	0.177	0.242	0.239	0.048

根据上表即可得出第一类数据的回归模型，如下所示：

$$y = 0.0512x_1 + 0.003x_4 + 0.061x_5 + 0.179x_6 + 0.177x_8 + 0.242x_9 + 0.239x_{10} + 0.048x_{11}$$

2、第二类

我们可以得到五个主成分，主成分可由 x_2、x_3、x_4、x_5、x_6、x_7、x_8、x_9 和 x_{10} 决定，因此我们可以得到 y 关于 x_2、x_3、x_4、x_5、x_6、x_7、x_8、x_9 和 x_{10} 的主成分回归函数。根据前文所提出的权重确定方法，所得到的各变量的权重值如下表所示。

表 2.16 第二类变量所对应的权重值

变量	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
权重	-0.024	0.111	-0.081	0.152	0.264	0.084	0.283	0.051	0.115

根据上表即可得出第二类数据的回归模型，如下所示：

$$y = -0.024x_2 + 0.111x_3 - 0.081x_4 + 0.152x_5 + 0.264x_6 + 0.084x_7 + 0.283x_8 + 0.051x_9 + 0.115x_{10}$$

3、第三类

我们可以得到六个主成分，主成分可由 x_1、x_2、x_3、x_4、x_5、x_6、x_8、x_9、x_{10} 和 x_{11} 决定，因此我们可以得到 y 关于 x_1、x_2、x_3、x_4、x_5、x_6、x_8、x_9、x_{10} 和 x_{11} 的主成分回归函数。根据前文所提出的权重确定方法，所得到的各变量的权重值如下表所示。

表 2.17 第三类变量所对应的权重值

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_8	x_9	x_{10}	x_{11}
权重	0.14	-0.011	0.121	-0.156	0.141	0.167	0.193	0.108	0.12	0.177

根据上表即可得出第三类数据的回归模型，如下所示：

$$y = 0.14x_1 - 0.011x_2 + 0.121x_3 - 0.156x_4 + 0.141x_5 + 0.167x_6 + 0.193x_8 + 0.108x_9 + 0.12x_{10} + 0.177x_{11}$$

4、第四类

我们可以得到五个主成分，主成分可由 x_1、x_2、x_4、x_6、x_8、x_9 和 x_{10} 决定，因此我们可以得到 y 关于 x_1、x_2、x_4、x_6、x_8、x_9 和 x_{10} 的主成分回归函数。根据前文所提出的权重确定方法，所得到的各变量的权重值如下表所示。

表 2.18 第四类变量所对应的权重值

变量	x_1	x_2	x_4	x_6	x_8	x_9	x_{10}
权值	-0.093	0.099	0.138	0.294	0.293	0.166	0.142

根据上表即可得出第四类数据的回归模型，如下所示：

$$y = -0.093x_1 + 0.099x_2 + 0.138x_4 + 0.294x_6$$
$$+0.293x_8 + 0.166x_9 + 0.142x_{10}$$

5、第五类

我们可以得到五个主成分，主成分可由 x_1、x_2、x_3、x_4、x_5、x_6、x_8、x_9 和 x_{10} 决定，因此我们可以得到 y 关于 x_1、x_2、x_3、x_4、x_5、x_6、x_8、x_9 和 x_{10} 的主成分回归函数。根据前文所提出的权重确定方法，所得到的各变量的权重值如下表所示。

表 2.19 第五类变量所对应的权重值

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_8	x_9	x_{10}
权重	0.127	0.091	0.059	-0.11	0.11	0.222	0.227	0.123	0.128

根据上表即可得出第一类数据的回归模型，如下所示：

$$y = 0.127x_1 + 0.091x_2 + 0.059x_3 - 0.11x_4 + 0.11x_5$$
$$+0.222x_6 + 0.227x_8 + 0.123x_9 + 0.128x_{10}$$

2.5 量化分级

根据前面的聚类分析结果，我们将附件 1 中的恐怖袭击事件按照地区分成了五类，随后对这五类进行了因子分析，分别确定了对应的主成分回归模型。由于恐怖袭击事件的危害程度在不同地区有不同的评判指标，因此不同分类的主成分回归模型不同。

本节中我们需要对不同的恐怖袭击事件按照危害程度分为一至五级，级别分别定义为特别严重、较严重、严重、一般和不严重五个等级。我们首先确定某一恐怖袭击事件的类别，然后根据该类别的主成分回归模型计算目标函数，然后根据量化标准判断该事件的危害等级。危害等级量化标准如下：

1、第一类

我们根据前面建立的回归模型计算第一类中每一个事件的目标值，最大值为2294，最小值为 0.0512。因此所建立的量化标准为：

特别严重：目标函数>500

较严重：200<目标函数<500

严重：50<目标函数<200

一般：10<目标函数<50

不严重：目标函数<10

2、第二类

我们根据前面建立的回归模型计算第二类中每一个事件的目标值，最大值为423.72，最小值为-0.024。因此所建立的量化标准为：

特别严重：目标函数>200

较严重：100<目标函数<200

严重：25<目标函数<100

一般：10<目标函数<25

不严重：目标函数<10

3、第三类

我们根据前面建立的回归模型计算第三类中每一个事件的目标值，最大值为121.29，最小值为 0.094。因此所建立的量化标准为：

特别严重：目标函数>100

较严重：50<目标函数<100

严重：25<目标函数<50

一般：10<目标函数<25

不严重：目标函数<10

4、第四类

我们根据前面建立的回归模型计算第四类中每一个事件的目标值，最大值为63.08，最小值为 0.006。因此所建立的量化标准为：

特别严重：目标函数>50

较严重：25<目标函数<50

严重：15<目标函数<25

一般：5<目标函数<15

不严重：目标函数<5

5、第五类

我们根据前面建立的回归模型计算第四类中每一个事件的目标值，最大值为542.75，最小值为 0.15。因此所建立的量化标准为：

特别严重：目标函数>150

较严重：75<目标函数<150

严重：30<目标函数<75

一般：10<目标函数<30

不严重：目标函数<10

2.6 十大恐怖袭击事件

根据建立的量化分级标准，我们对附件一中的恐怖袭击事件进行分级，最终可以得到十大恐怖袭击事件如下：

1、恐怖袭击事件编号为 200109110004，目标函数为 2294；

2、恐怖袭击事件编号：199808070002，目标函数为 542；
3、恐怖袭击事件编号：201608010021，目标函数为 423；
4、恐怖袭击事件编号：200409010002，目标函数为 259；
5、恐怖袭击事件编号：201710010018，目标函数为 220；
6、恐怖袭击事件编号：201603080001，目标函数为 177；
7、恐怖袭击事件编号：200802030004，目标函数为 151；
8、恐怖袭击事件编号：200403110003，目标函数为 127；
9、恐怖袭击事件编号：201607140001，目标函数为 127；
10、恐怖袭击事件编号：200607120004，目标函数为 121；

2.7 事件评级

1、200108110012

此恐怖袭击事件属于第五类(SSA)，我们将各量化指标代入第五类所对应的主成分回归方程，最终得到目标函数值为 56，等级为第三级(严重)。

2、200511180002

此恐怖袭击事件属于第二类(MENA)，我们将各量化指标代入第二类所对应的主成分回归方程，最终得到目标函数值为 18，等级为第四级(一般)。

3、200901170021

此恐怖袭击事件属于第五类(SSA)，我们将各量化指标代入第五类所对应的主成分回归方程，最终得到目标函数值为 56，等级为第三级(严重)。

4、201402110015

此恐怖袭击事件属于第二类(MENA)，我们将各量化指标代入第二类所对应的主成分回归方程，最终得到目标函数值为 3.95，等级为第五级(不严重)。

5、201405010071

此恐怖袭击事件属于第五类(SSA)，我们将各量化指标代入第五类所对应的主成分回归方程，最终得到目标函数值为 7.3，等级为第五级(不严重)。

6、201411070002

此恐怖袭击事件属于第三类(SAS)，我们将各量化指标代入第三类所对应的主成分回归方程，最终得到目标函数值为 3.5，等级为第五级(不严重)。

7、201412160041

此恐怖袭击事件属于第二类(MENA)，我们将各量化指标代入第二类所对应的主成分回归方程，最终得到目标函数值为 8.6，等级为第五级(不严重)。

8、201508010015

此恐怖袭击事件属于第二类(MENA)，我们将各量化指标代入第二类所对应的主成分回归方程，最终得到目标函数值为 0.32，等级为第五级(不严重)。

9、201705080012

此恐怖袭击事件属于第三类(SAS)，我们将各量化指标代入第三类所对应的主成分回归方程，最终得到目标函数值为 10.7，等级为第四级(一般)。

表 2.20 典型事件危害级别

事件编号	危害级别
200108110012	第三级(严重)
200511180002	第四级(一般)

200901170021	第三级(严重)
201402110015	第五级(不严重)
201405010071	第五级(不严重)
201411070002	第五级(不严重)
201412160041	第五级(不严重)
201508010015	第五级(不严重)
201705080012	第四级(一般)

第三章 任务二解答

根据已知犯罪者的样本数据去预测未知犯罪者的样本数据，可以分为两步，首先利用犯罪者的样本数据进行分类训练，得到样本特征—犯罪者的模型，其次利用该模型对未知犯罪者的事件进行预测。由于所分类的维数较多，并且数据量较大，考虑训练效率与模型精度两个指标分类树[5]是较好的选择。

在进行训练之前首先需要对 1998-2016 年的数据进行筛选，得到较为理想的数据。其次，在诸多特征中找出较为合理的特征，用以训练分类树。最后，根据得到的分类树模型对待预测的数据进行预测，在结果中筛选出危害程度较高的前五组组织，进而用这五组组织对 17 年的 10 典型数据进行嫌疑程度排序。

3.1 数据分析

3.1.1 数据筛选

在统计学分析中，早期的数据对现在或今后形式的预测可能不具有指导意义，但是对已有的数据进行统计分析看出，较多犯罪组织的犯罪特征随时间变化不大，因此早期的数据对现在依旧具有指导意义，统计 1998-2016 年的数据进行训练是合理的。

在 1998-2016 年所有样本中，首先筛选出所有具有明确组织或个人的事件，总共得到 36457 组数据。进而对该样本按照 gname（组织或个人）进行统计，可以得到 1273 组统计结果，每一个组织或个人所犯事件从 1-5519 不等。

对该样本按照 gname 进行观察分析可以看出，一些组织在某一年份后再未有过犯罪事件；一些组织犯罪事件发生的年份跨度时间较短；一些组织犯罪事件发生的年份具有跳跃性。这几种情况均会对分类模型的精确度产生影响，因此首先需要根据犯罪组织的时间跨度对数据进行筛选。

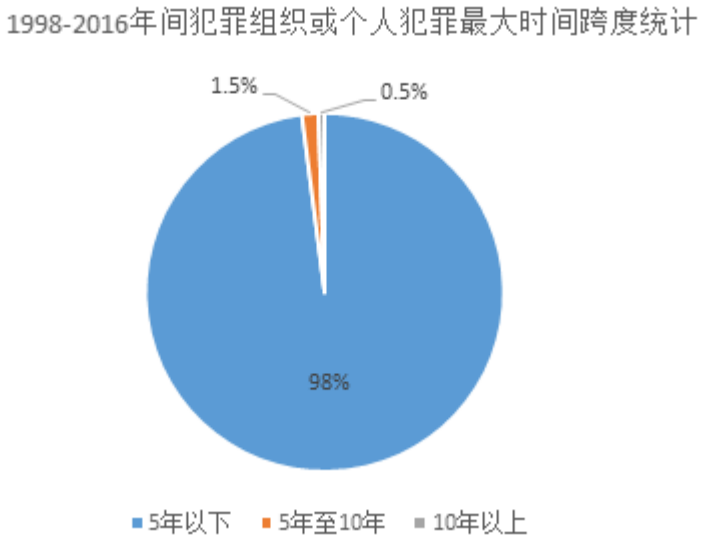

图 3.1 犯罪组织或个人最大时间跨度统计图

图 3.1 统计近 19 年出现的 1274 组犯罪组织或个人的最大犯罪时间跨度，（如

对于 AI-Qaidn 在其存在犯罪记录的 1998-2011 期间，2005-2010 年间没有犯罪事件发生，其他年间均有犯罪记录或间隔小于 5 年，则其最大犯罪时间跨度为 5 年），1248 组组织的时间跨度小于 5 年，有 19 组的犯罪事件跨度在 5 年与 10 年之间，仅有 7 组的犯罪事件跨度超过 10 年，其比例为 0.5%，为小概率事件。因此，可以做出如下合理推断：

近十年中没有出现犯罪记录的组织或个人不再存在。

根据该推断忽略 2007 年后（不含 2007 年）再未出现犯罪记录的组织所对应的事件，将对应的样本数据剔除。

在原有的 1274 组犯罪组织中，犯罪次数统计如下图：

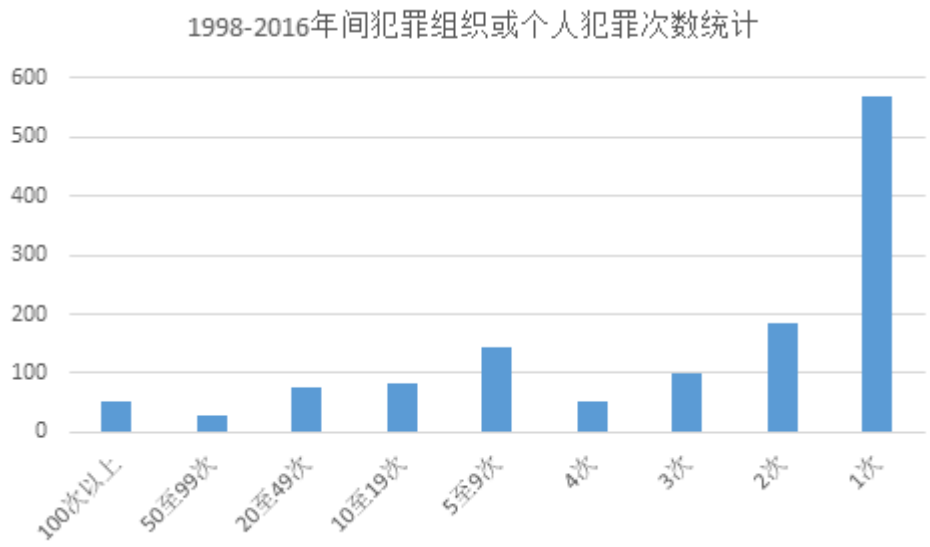

图 3.2 犯罪组织或个人犯罪次数统计图

通过图 3.2 可以看出仅出现过一次的犯罪组织或个人达到 567 组，占总数的 44.51%，即仅出现 1 次的组织接近总组织数的一半，如果把这些数据均统计在内，则必然对分类结果产生较大干扰，因此有必要对出现 1 次的组织或个人进行筛选。根据之前对时间跨度的统计可以看出，攻击时间跨度大于等于 5 年的组织仅占 2%，并且每个组织时间跨度较大近会出现 1 次，如果以事件总数 36457 作为基数，则攻击时间跨度大于等于 5 年的概率可以忽略不计。因此对仅发生 1 次攻击时间的组织进行如下合理推断：

1998-2016 年间攻击次数等于一次的组织，攻击时间在 2010 年之前的不会再发动攻击。

根据该假设在攻击次数等于 1 次的组织中，剔除掉攻击时间早于 2010 年的样本数据。最终筛选得到 35446 组样本数据，849 组犯罪组织或个人。

3.1.2 特征选择

在所有样本数据中包含有 40 多组特征，需要对这些特征进行筛选。其中有些特征信息不完全，即有些样本数据没有该特征（如 motive、及几乎所有的 type2-n），这样的特征会给分类树的训练带来不变，并且也不适合作为分类的标准；有些特征难以量化，并且价值含量低，（如 latitude、longitude，该项特征完全可以用地区与国家进行代替）；有些特征较为简单，无法区分目标，（如 suiside，几乎绝大部分的样本数据的该项特征均是 0）；有些特征为连续变化的数值，（如

nkill、nwound）并且相同组织或个人造成的事件波动范围较大，即该特征对目标的预测意义不大。

经过简单筛选得到初步候选的特征，包括 crit1（入选标准 1），crit2（入选标准 2），crit3（入选标准 3），region（地区），country（国家），attacktype1（攻击类型），weapontype1（武器类型 1），targtype1（目标类型 1）共 8 个特征。

一般特征选择的依据是该项特征具有较高的信息增益，即该特征可以将数据集中的大部分分割成子集，进而可以通过其他分类特征确定属于哪个组织。但是对于不同问题信息增益阈值的选择也不尽相同，并且由于在特征数量较多时，计算量过于复杂，因此本文随机选择十六组数据，对初步筛选的特征进行统计分析来选取特征，统计结果如表 3.1 所示。

其中，gname1-16 分别表示 Taliban、Islamic State of Iraq and the Levant (ISIL)、Al-Shabaab、Kurdistan Workers' Party (PKK)、Al-Qaida in Iraq、Al-Qaida in Iraq、Muslim extremists、Hamas (Islamic Resistance Movement)、Allied Democratic Forces (ADF)、Khorasan Chapter of the Islamic State、Adan-Abyan Province of the Islamic State、Hutu extremists、Tawhid and Jihad、Bangladesh Nationalist Party (BNP)、Oglaigh na hEireann、Democratic Front for the Liberation of Palestine、Al-Qaida in the Islamic Maghreb (AQIM)。

表 3.1 不同组织的特征数量统计

Gname	region	country	attacktype1	targtype1	weaptype1	crit1	crit2	crit3
1	2	3	9	20	10	1	2	2
2	1	6	8	19	9	1	2	2
3	1	6	9	20	6	2	2	2
4	2	8	8	16	6	1	1	2
5	1	2	6	16	3	1	1	2
6	8	40	9	16	8	2	2	2
7	2	3	6	16	7	2	2	2
8	1	2	7	11	5	1	1	2
9	1	2	6	13	5	1	1	2
10	1	1	5	8	3	1	2	2
11	2	3	4	8	5	1	1	2
12	1	2	6	4	10	1	2	2
13	1	1	3	9	3	1	1	2
14	1	1	5	5	3	1	2	2
15	1	2	2	3	2	1	1	2
16	2	9	6	15	5	1	1	2

从表 3.1 可以推测大部分犯罪组织或个人的犯罪区域或国家较为集中，某一组织犯罪事件的攻击类型、武器类型、目标类型数量虽然与事件总数呈正相关，但考虑到绝大部分组织的犯罪事件总数较少，完全可以选择该特征进行训练。此外，几乎所有事件的 ctri3 有两种情况，并且有较多事件的 ctri2 有两种情况，这使得利用该特征难以对事件进行预测，与之相反，虽然大部分事件的 ctri1 具有一个明确的值，但是对于相同组织，但是几乎所有的组织的该特征为同一值，因此 ctri1-3 不具有较好的分类特性。

综上分析，选择 region（地区），country（国家），attacktype1（攻击类型），weapontype1（武器类型 1），targtype1（目标类型 1）5 个特征作为训练特征。

3.2 模型构建

3.2.1 模型选择

本章开篇分析采用分类树是较好的选择，在众多的分类树算法中，CART（Classification And Regression Tree）通过计算 GINI 值选择分裂的属性，

分类树下面有两个关键的思想。第一个是关于递归地划分自变量空间的想法；第二个想法是用验证数据进行剪枝。

$$L=(X_1,X_2,...,X_m;Y)$$

其中 X 为样本数据，m 为样本总数，Y 为训练标签。

$$X_n=(x_n^1,x_n^2,...,x_n^k)$$

其中，k 为特征维数

$$Y=(y_1,y_2,...,y_l)$$

其中 l 为训练标签的维数

$$GINI(S)=1-\sum_{i=1}^{m}p_i^2$$

p_i 表示分类结果中第 i 个类别出现的频率

$$Gain(x_n^i)=\frac{n_1}{n}GINI(x_n^i)+\frac{n-n_1}{n}GINI(\overline{x_n^i})$$

对于 x_n，选择增益最小的 x_n^i 作为该特征的二分类属性

$$SplitGain(x_n)=\min_{x_n^i\in x_n}(Gain(x_n^i))$$

对于样本 X，遍历所有的特征 $x_n^1,x_n^2,...,x_n^k$，选择增益最小者作为整个样本集的分类属性。

$$SplitGain(X)=\min_{x_n\in X}(\min_{x_n^i\in x_n}(Gain(x_n^i)))$$

综上所述，本文采用的分类算法流程如下所示：

利用所有样本创建书的根节点 Root

If 当前节点中的 Samples 都属于同一类或者分裂特征为空

　　标记此叶点为叶子节点

　　返回此节点树 Root

Else

　　按照选择标准选择最好的分裂特征进行分裂

　　For 每一个子节点

　　从 Samples 中选择子集，作为此节点的样本创建子树

　　End

End

对于决策树还需要进行剪枝处理。令决策树的非叶子节点为$\{N_1, N_2, ..., N_n\}$

计算所有非叶子节点的表面误差率增益值$\alpha = \{\alpha_1, \alpha_2, ..., \alpha_n\}$

选择表面误差率增益值α_i最小的非叶子节点N_i（若多个非叶子节点具有相同小的表面误差率增益值，选择节点数最多的非叶子节点），对N_i进行剪枝

表面误差率增益值的计算公式：$\alpha = \dfrac{R(t) - R(T)}{N(T) - 1}$

$$R(t) = r(t) \cdot p(t)$$

其中$R(t)$表示叶子结点的误差代价，$r(t)$为节点的错误率，$p(t)$为节点数据量的占比。$R(T)$表示数的误差代价，

$$R(T) = \sum_{i}^{m} r_i(t) \cdot p_i(t)$$

$r_i(t)$表示节点 i 的错误率，$p_i(t)$表示节点 i 的数据节点占比，$N(T)$表示子树节点个数。

3.2.2 模型训练

利用 Matlab 对样本数据进行训练，训练过程中为了提高模型的泛化能力，采用 K-折交叉验证法，即将全部训练集 X 分成 K 个子集，每一次训练不重复地选择其中一个作为测试集，剩余的 K-1 个数据作为训练集，重复训练直到所有的子

集均被用于测试，最终可以得到泛化能力最优的模型。本文算法采用 5-折交叉验证法，最终训练准确率达到 86.6%。

3.2.3 模型验证

本文为了验证模型的精度，还采用了 2017 年已确定犯罪者的样本数据进行验证。在整理出的 6126 组数据中模型分类准确率达到 79.4%，如果排除掉新兴的组织存在的必然误差外（即无法预测 2017 年首次出现的组织），分类准确率会更高。

3.3 预测

在验证了模型的精度后，需要完成对 2015-2016 年未知组织或个人的数据的预测。在样本集中提取 2015-2016 年未知组织或个人的数据，共有 12366 组样本数据，利用上述模型对其进行预测，最终得到 41 个组织。

3.4 嫌疑度排序

3.4.1 组织危险程度排序

组织的危险程度取决于其造成的危害，因此利用 nkill、nwound、property 作为危险程度的排列特征，此外 success、发生次数两个特征与组织的危险程度相关，因此取特征$\{x_1, x_2, x_3, x_4, x_5\}$分别为:nkill nwound property success，gname，首先将 41 个组织对这五个特征进行统计得到表 3.2。其中，label 为组织的代号，其对应关系如附表一。

表 3.2 反映组织危害程度特征表

label	nkill	nwound	property	success	gname 次数
1	3.86	3.92	-1.08	0.91	5519
2	8.19	7.31	-1.77	0.88	3994
3	3.39	3.21	-2.16	0.92	2032
4	9.79	5.57	-0.85	0.93	1477
5	1.03	0.93	-0.26	0.8	1211
6	0.83	0.96	-0.46	0.87	1111
7	1.35	3.16	-0.95	0.92	977
8	3.4	4.27	-2.96	0.86	884
9	4.58	7.35	-0.68	0.94	797
11	1.86	2.53	-0.49	0.89	720
12	3.61	3.36	-1.46	0.88	644
13	2.87	2.72	-3.46	0.93	599

16	0.57	1.23	-3.11	0.66	366
19	3.12	2.69	-1.16	0.88	316
25	4.7	8.36	-2.03	0.95	262
27	0.98	2.72	-0.66	0.95	239
30	3.05	2.71	0.32	0.99	195
31	4.86	9.46	-1.15	0.87	183
33	6.26	2.06	-0.8	0.94	163
37	11.22	2.62	-1.94	0.97	153
43	2.85	2.16	-2.43	0.94	135
47	1.04	2.17	-0.72	0.85	109
50	0.02	0.2	0.91	0.97	105
56	11.58	7.76	-0.39	0.98	81
58	0.05	1.65	-3.02	0.94	80
65	3.37	1.26	-0.85	0.99	65
66	1.48	0.68	-0.38	1	61
70	23	13.3	-1.61	0.93	58
71	7.71	14.7	-1.41	0.85	57
73	10.83	5.44	-3.69	0.91	57
75	0	0.01	0.86	0.88	56
79	2.24	3.17	-1.26	0.98	51
80	7.6	3.26	-2.4	0.96	51
83	1.12	1.71	-1.05	0.87	49
88	0.12	0.04	0.97	1	43
113	10.67	14.06	-1.5	0.94	30
163	17.21	6.87	-0.29	0.92	18
198	4.41	2.28	-2	1	12
209	0.28	0.64	0.85	1	11
213	0.5	2.3	-2.6	0.99	10
262	0	0	1	1	7

根据表 3.2，采用如下公式来计算危险程度：

$$sum = \sum_{i=1}^{4} \frac{x_i}{\max(x_i)} + 2\frac{x_5}{\max(x_5)}$$

其中 $\frac{x_i}{\max(x_i)}$ 表示该列数据除以该列中的最大值，即归一化，$2\frac{x_5}{\max(x_5)}$ 表示特征 x_5 的权值较大，计算结果如表 3.3，选取 sum 最大的前五个组织作为危险程度较高的组织，并对其进行排序得到表 3.4。其中 label 为组织的代号，其对应关系见附表一。

表 3.3 危害程度特征归一化表

label	nkill	nwound	property	success	gname	sum
2	0.475886	0.497279	0.479675	0.88	1.447364	3.780203
1	0.224288	0.266667	0.292683	0.91	2	3.693638
70	1.336432	0.904762	0.436314	0.93	0.021018	3.628527
113	0.619988	0.956463	0.406504	0.94	0.010872	2.933827
73	0.629285	0.370068	1	0.91	0.020656	2.930009
71	0.447995	1	0.382114	0.85	0.020656	2.700765
262	0	0	-0.271	1	0.002537	2.693638
3	0.196979	0.218367	0.585366	0.92	0.736365	2.657077
4	0.568855	0.378912	0.230352	0.93	0.535242	2.643361
163	1	0.467347	0.078591	0.92	0.006523	2.472461
8	0.19756	0.290476	0.802168	0.86	0.320348	2.470552
25	0.273097	0.568707	0.550136	0.95	0.094945	2.436885
13	0.166764	0.185034	0.937669	0.93	0.217068	2.436535
37	0.651947	0.178231	0.525745	0.97	0.055445	2.381368
56	0.672865	0.527891	0.105691	0.98	0.029353	2.3158
80	0.441604	0.221769	0.650407	0.96	0.018482	2.292261
9	0.266124	0.5	0.184282	0.94	0.28882	2.179227
31	0.282394	0.643537	0.311653	0.87	0.066316	2.173901
43	0.165601	0.146939	0.658537	0.94	0.048922	1.959999
198	0.256246	0.155102	0.542005	1	0.004349	1.957702
12	0.209762	0.228571	0.395664	0.88	0.233376	1.947373
58	0.002905	0.112245	0.818428	0.94	0.028991	1.902569

213	0.029053	0.156463	0.704607	0.99	0.003624	1.883746
7	0.078443	0.214966	0.257453	0.92	0.35405	1.824911
16	0.03312	0.083673	0.842818	0.66	0.132633	1.752245
33	0.363742	0.140136	0.216802	0.94	0.059069	1.719749
79	0.130157	0.215646	0.341463	0.98	0.018482	1.685748
19	0.18129	0.182993	0.314363	0.88	0.114513	1.67316
11	0.108077	0.172109	0.132791	0.89	0.260917	1.563894
65	0.195816	0.085714	0.230352	0.99	0.023555	1.525438
6	0.048228	0.065306	0.124661	0.87	0.402609	1.510804
27	0.056944	0.185034	0.178862	0.95	0.08661	1.457449
5	0.059849	0.063265	0.070461	0.8	0.438848	1.432423
83	0.065078	0.116327	0.284553	0.87	0.017757	1.353715
30	0.177223	0.184354	-0.08672	0.99	0.070665	1.33552
47	0.06043	0.147619	0.195122	0.85	0.0395	1.292671
66	0.085997	0.046259	0.102981	1	0.022105	1.257342
209	0.01627	0.043537	-0.23035	1	0.003986	0.833441
50	0.001162	0.013605	-0.24661	0.97	0.03805	0.776205
88	0.006973	0.002721	-0.26287	1	0.015583	0.762404
75	0	0.00068	-0.23306	0.88	0.020294	0.667911

对应危险程度最高的恐怖组织如表 3.4 所示。

表 3.4 前五个组织作为危险程度较高的组织

组织	危险等级
Islamic State of Iraq and the Levant (ISIL)	1
Taliban	2
Sudan People's Liberation Movement – North	3
Uighur Separatists	4
Sudan People's Liberation Movement in Opposition (SPLM-IO)	5

3.4.2 嫌疑人嫌疑度排序

首先统计表 2 中的十组恐怖袭击事件的特征，如表 3.5 所示。

表 3.5 恐怖袭击事件特征统计

eventid	iyear	country	region	attacktype1	targtype1	weaptype1
201701090031	2017	95	10	3	14	6
201702210037	2017	4	6	9	3	13
201703120023	2017	1004	11	7	4	8
201705050009	2017	1004	11	2	19	5
201705050010	2017	1004	11	2	19	5
201707010028	2017	41	11	9	14	13
201707020006	2017	200	10	3	4	6
201708110018	2017	4	6	3	14	6
201711010006	2017	4	6	3	19	6
201712010003	2017	95	10	3	14	6

随后统计前面得出的排名前五的恐怖组织的特征，如表 3.6 所示。

表 3.6 排名前五的恐怖组织的特征

组织	country	region	attacktype1	targettype1	weaptype1
1	18,21,60,69,74,75,93,95,102,110,113,121,155,160,167,173,182,200,208,209,228	5,7,8,9,10,11	1,2,3,4,5,6,7,9	1,2,3,4,6,7,8,9,10,12,13,14,15,17,18,19,20	2,5,6,8,9,10,11,12,13
2	4,153,210	6,7	1,2,3,4,5,6,7,8,9	1,2,3,4,6,7,8,9,10,12,13,14,15,16,17,18,19,20,21,22	2,3,5,6,8,9,10,11,12,13
3	195	11	2,3,6,7,9	1,2,4,7,8,12,14	5,6,13
4	44	4	1,2,3,4,6,8,9	1,2,3,4,6,14,15,19	6,8,9,13
5	1004	11	1,2,3,6,7,9	1,2,3,4,7,14,15,19,21	5,6,8,13

采用如下的方法进行嫌疑度排序：

1、首先考虑 region 特征。对每一组未知数据通过 region 选择可能的犯罪组织，如果 region 不相同则不考虑该组织。

2、其次考虑 country 特征。对于选出的组织，按照国家进行筛选，出现相同国家的组织优先。

3、继而考虑 attacktype1、targettype1、weaptype1 特征。如果国家与组织均相同，则按照 attacktype1、targettype1、weaptype1 特征进行筛选，特征相同次数较多的组织优先。

4、最后，如果出现上述条件均相同的组织，则按照 attacktype1、targettype1、weaptype1 特征数较少的组织确定嫌疑度。

如:对于样本数据 201703120023，首先其 region 特征为 11，根据 region 筛选出组织 1,3,5，其余组织不再考虑。其次根据 country 得到嫌疑度最高者为 5，1、3 两个组织嫌疑度较低。最后，样本数据的后三个特征为 7,4,8，组织 1 与之相重 3 个，组织 3 与之相重 2 个，因此组织 1 的嫌疑度较高。综上所述，其嫌疑度排序为 5,1,3。

同理可以得到嫌疑度排序如下表。

表 3.7 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
样例 XX	4	3	1	2	5
201701090031	1	—	—	—	—
201702210037	2	—	—	—	—
201703120023	5	1	3	—	—
201705050009	5	1	3	—	—
201705050010	5	1	3	—	—
201707010028	1	3	5	—	—
201707020006	1	—	—	—	—
201708110018	2	—	—	—	—
201711010006	2	—	—	—	—
201712010003	1	—	—	—	—

第四章 任务三解答

在本节中，我们首先对附件 1 中的原始数据进行了预处理，针对 2015 年~2017 年发生的恐怖袭击事件进行了统计分析，定性分析了近三年来恐怖袭击事件的特性及发展规律，从而在直观上有了相应的了解。恐怖袭击事件的指数受不同因素的影响，因此我们选取了最能客观反映恐怖袭击事件风险指标的因素。基于历史统计数据，利用多元统计分析中的因子分析建立了恐怖袭击风险的评估模型[6][7][8]，从而可以为 2018 年的反恐态势提出参考。

4.1 描述性统计特性分析

由于每个地区所面临的反恐态势都不尽一样，且不同地区由于经济、军事能力的差异，应对反恐攻击的能力也天差地别。因此，我们针对 12 个地区近三年所发生的恐怖袭击事件进行统计分析，从而可以直观地得到对应的发展规律。

(1) 恐怖活动热点地区分析

表 4.1 不同地区的受袭击次数

地区	2015 年	2016 年	2017 年	总袭击次数
NA	62	75	97	234
CAC	1	3	4	8
SA	176	159	172	507
EA	28	8	7	43
SEA	1072	1077	1020	3169
SAS	4585	3639	3430	11654
CA	10	17	7	34
WE	333	273	291	897
EE	683	134	110	927
MENA	6035	6115	3780	15930
SSA	1964	2077	1970	6011
AO	14	10	12	36
年总次数	14963	13587	10900	
年成功次数	12674	10975	8652	

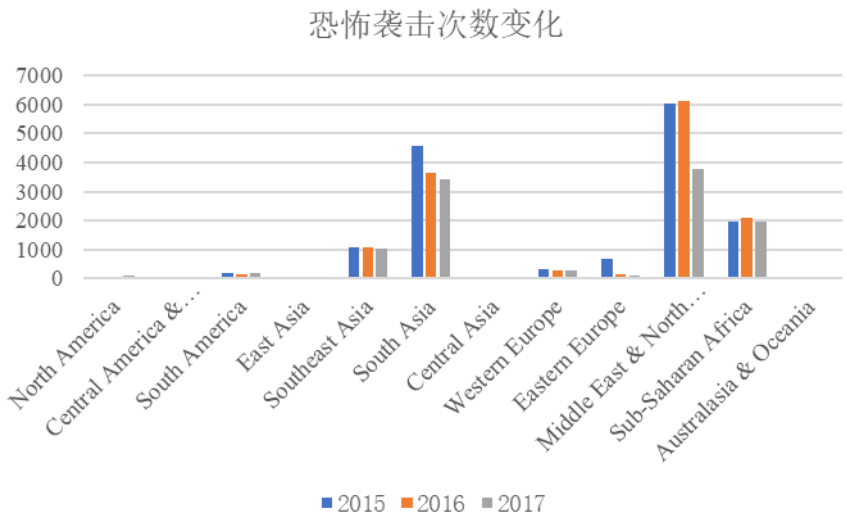

图 4.1 不同地区的恐怖袭击次数

从表 4.1 和图 4.1 可以看出，不同地区所遭受的恐怖袭击次数有着较大的差别。在南亚(SA)、中东(MENA)和非洲(SSA)地区，2015 年~2016 年它们分别遭受了 11654 次、15930 次和 3011 次恐怖袭击，共占总袭击次数的 85%，可见这几个区域是全球恐怖袭击事件最为频发的地点，反恐形势较为严峻。2015 年~2017 年中，非洲地区的受袭击次数大体保持一致；虽然南亚地区和中东地区的恐怖袭击次数有所减少，但仍然保持在较大的数量，因此这些地区的反恐部署工作仍然不可放松。

(2) 恐怖袭击后果分析

表 4.2 恐怖袭击事件伤亡人数

	死亡人数			受伤人数		
地区	2015 年	2016 年	2017 年	2015 年	2016 年	2017 年
NA	62	73	124	70	173	963
CAC	0	9	4	0	0	1
SA	128	87	101	145	179	156
EA	123	32	16	84	44	77
SEA	696	589	811	1452	1154	1008
SAS	8300	7803	7663	10233	9277	8990
CA	13	21	6	12	25	7
WE	171	238	83	526	900	509
EE	790	112	101	1432	107	183

MENA	17852	19205	10819	23752	23433	8392
SSA	10762	6649	6712	6397	4642	4617
AO	2	0	4	0	1	24

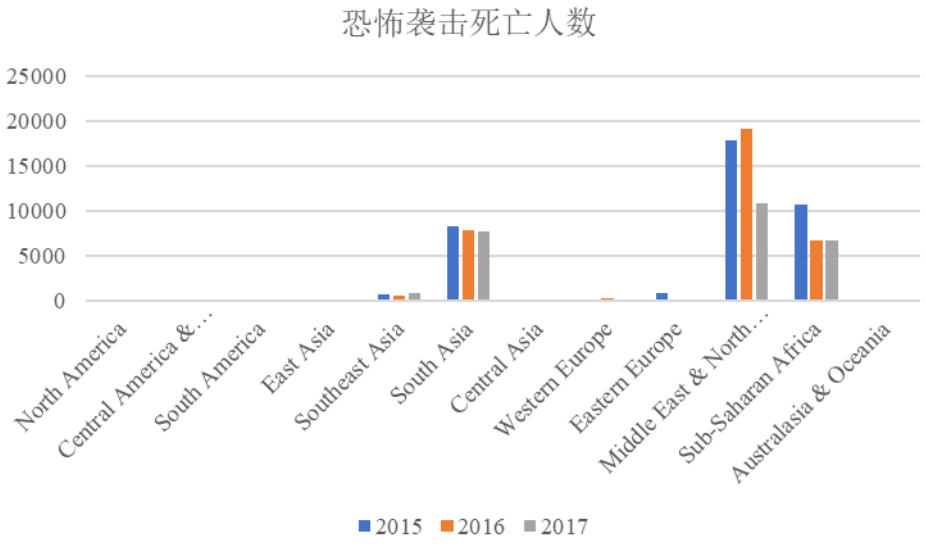

图 4.2 恐怖袭击中死亡人数

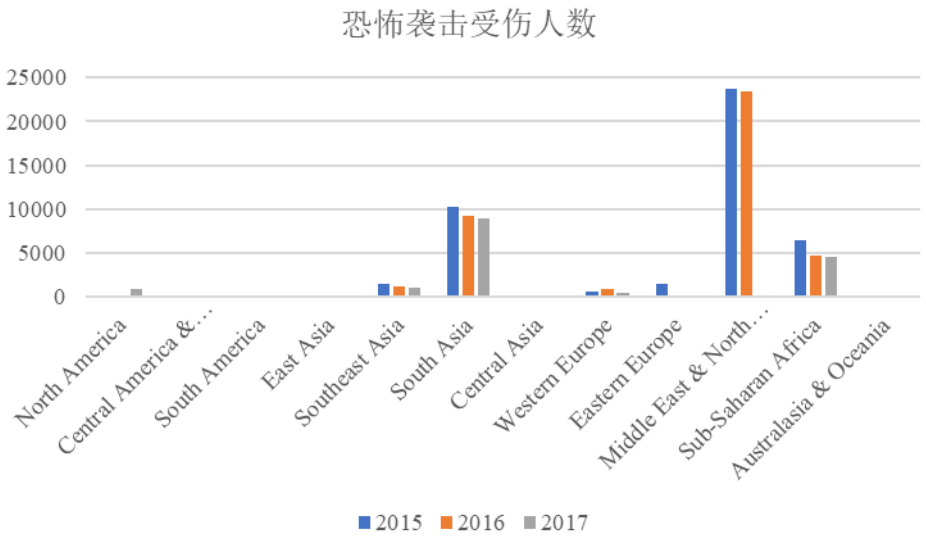

图 4.3 恐怖袭击中受伤人数

从表 4.2、图 4.2 和图 4.3 可以看出，2015 年至 2017 年间，全球每年因恐怖袭击事件而受伤、死亡的人数至少会达到 50000。特别是在南亚、中东和非洲这三个地区，它们的伤亡人数占到了全球伤亡人数的 90%以上，恐怖主义十分猖獗，因此对这三个地区加强反恐是十分有必要的。

(3) 恐怖活动热点国家分析

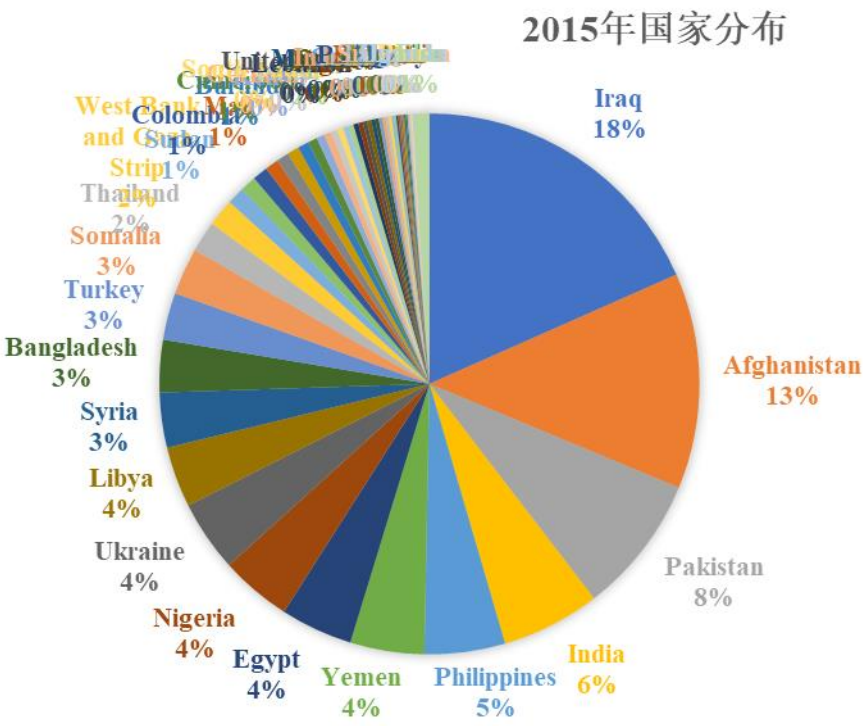

图 4.4 2015 年恐怖活动热点国家

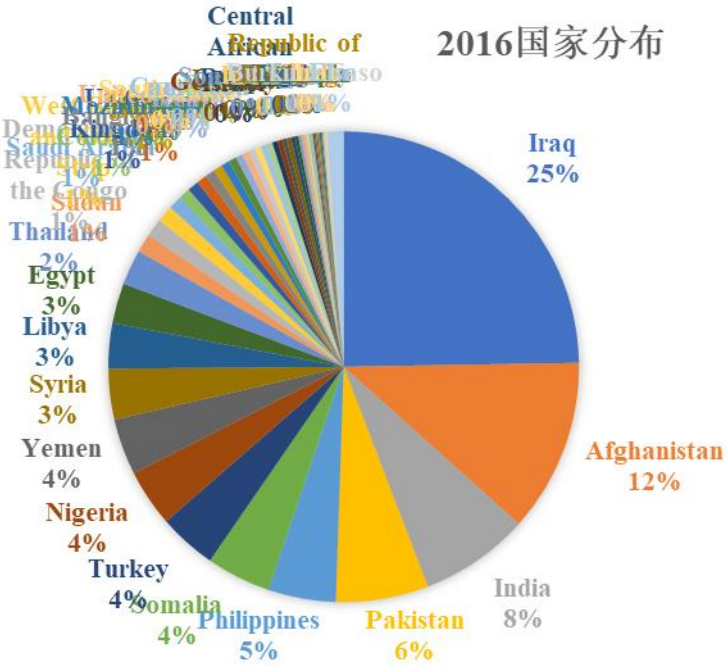

图 4.5 2016 年恐怖活动热点国家

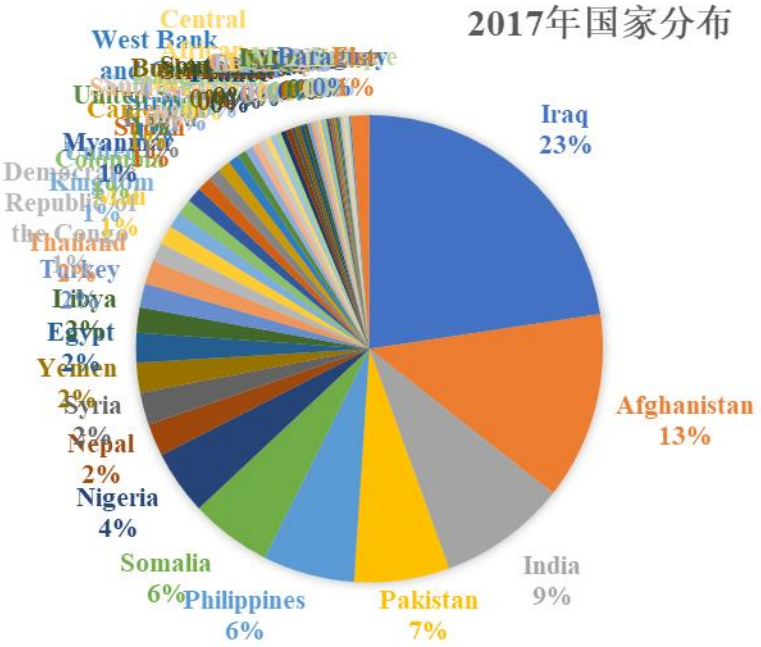

图 4.6 2017 年恐怖活动热点国家

从上图我们可以看出，世界上恐怖袭击活动最为猖獗的十个国家分别为：伊朗、阿富汗、印度、巴基斯坦、菲律宾、索马里、尼日利亚、尼泊尔、叙利亚和也门。它们境内发生的恐怖袭击事件约占全球恐怖袭击时间中的 74%。这些国家绝大部分分布在中东、非洲和南亚地区。因此对于袭击活动不仅仅要关注地区，而且对这些恐怖活动频繁的国家也应该引起重视，加强反恐活动。

4.2 选取反映恐怖袭击活动的指标

恐怖袭击活动的危害性受不同指标的影响，因此在我们分析恐怖活动风险指数前，应确定最能反映该指数的关键评价指标。在本文中，我们选取了六个风险评价指标：

1、 每年恐怖袭击事件发生的总次数 S_1

2、 每年恐怖袭击事件的总成功次数 S_2

3、 每年恐怖袭击事件导致的总死亡人数 S_3

4、 每年恐怖袭击事件所导致的总受伤人数 S_4

5、 每年导致人员死亡的恐怖袭击次数 S_5

6、 每年恐怖袭击所导致的财产损失 S_6

4.3 恐怖袭击事件风险评估模型

上面选取的风险评价指标最直接反映了某一具体恐怖袭击事件风险指数，并且每一个指标的影响程度不尽相同 (即我们需要对这些指标赋予不同的权值)。如果权值的选取具有很大的主观性，这将导致风险评估模型无法正确地反映实际情况，因此我们采用因子分析方法来选取需要的权值，该方法不需要主观确定各

项风险评价指标的权重，而是根据样本的实际记录值自动得到权重，这样可以消除主观因素，为政府提供客观的评价结果。恐怖事件风险评估模型如图 4.7 所示。

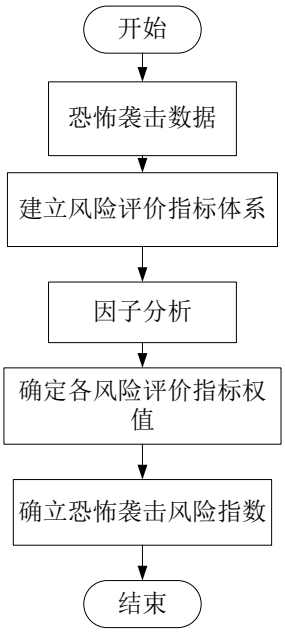

图 4.7 恐怖袭击事件风险评估模型流程图

4.4 因子分析及 KMO 检验

4.4.1 KMO（Kaiser-Meyer-Olkin)检验

我们在做因子分析之前，应对原始数据进行检验，判断其是否适合因子分析，通常采用的分析方法是 KMO 检验方法。

KMO 检验统计量通常用来对比变量之间的相关系数和偏相关系数，该统计量的取值范围为$[0,1]$。当 KMO 值越接近于 1 时，说明变量间的相关性越强，越适合做因子分析；相反，当 KMO 值越接近于 0 时，说明变量间的相关性非常弱。当 KMO 值大于 0.7 时，代表该原始数据适合做因子分析。下面分别对 2015 年、2016 年和 2017 年的原始数据进行 KMO 检验，检验结果如表 3.3-3.5 所示。

表 4.3 KMO 和巴特利特检验(2015 年)

KMO 取样适切性量数	.832	
巴特利特球形度检验	近似卡方	166.885
	自由度	15
	显著性	.000

表 4.4 KMO 和巴特利特检验(2016 年)

KMO 取样适切性量数	.719	
巴特利特球形度检验	近似卡方	186.021
	自由度	15
	显著性	.000

表 4.5 KMO 和巴特利特检验(2017 年)

KMO 取样适切性量数	.754	
巴特利特球形度检验	近似卡方	140.021
	自由度	15
	显著性	.000

从上表可知，我们对 2015 年、2016 年和 2017 年的数据进行 KMO 检验之后，可以分别得到对应的 KMO 系数为 0.832、0.719 和 0.754。说明变量间的相关性较强，原有变量适合做因子分析。

4.4.2 恐怖袭击事件因子分析结果

1、2015 年恐怖袭击事件分析

我们首先分析 2015 年的数据，对其进行因子分析。确定因子个数的原则是特征值大于 1。当特征值小于 1 时，表征该成分不具有可靠性，最终可将特征值大于 1 的提取出来作为主因子，相当于对整体数据进行了一定的降维处理。因子分析结果如表 4.6 和图 4.8 所示。

表 4.6 因子分析结果(2015 年)

	初始特征值			提取载荷平方和		
成分	总计	方差百分比	累积%	总计	方差百分比	累积%
1	5.224	87.060	87.060	5.224	87.060	87.060
2	0.651	10.853	97.913			
3	0.096	1.593	99.056			
4	0.028	0.470	99.976			
5	0.001	0.023	99.998			
6	0	0.002	1			

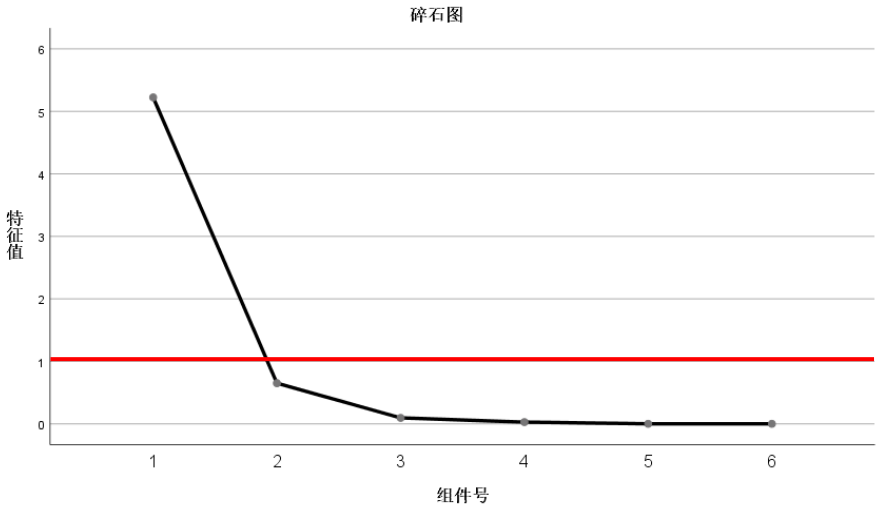

图 4.8 因子分析碎石图(2015 年)

根据上述的分析结果，我们可以建立 2015 年关于上述 6 个评价指标的恐怖袭击事件风险评估模型，最终可以得到恐怖事件的风险指数。根据我们对各指标的量化原则可以得到，风险指数越大，该地区发生恐怖袭击事件的风险越大；相反，风险指数越小则表示该地区发生恐怖袭击事件的风险越小。2015 的恐怖袭击事件风险评估模型如下所示：

$$Y = 0.178 * S_1 + 0.178 * S_2 + 0.176 * S_3 + 0.173 * S_4 + 0.179 * S_5 + 0.115 * S_6$$

2、2016 年恐怖袭击事件分析

我们首先分析 2016 年的数据，对其进行因子分析。因子分析结果如表 4.7 和图 4.9 所示。六个评价指标被分为了两个主因子。第一个主因子与五个指标相关联：每年恐怖袭击事件发生的总次数、每年恐怖袭击事件的总成功次数、每年恐怖袭击事件导致的总死亡人数、每年恐怖袭击事件所导致的总受伤人数和每年导致人员死亡的恐怖袭击次数。第二个主因子与盛夏的指标相关：每年恐怖袭击所导致的财产损失。

表 4.7 因子分析结果(2016 年)

	初始特征值			提取载荷平方和		
成分	总计	方差百分比	累积%	总计	方差百分比	累积%
1	4.957	82.613	82.613	4.957	82.613	82.613
2	1.004	16.737	99.350	1.004	16.737	99.350
3	0.029	0.487	99.837			
4	0.009	0.152	99.989			
5	0	0.007	99.996			
6	0	0.004	1			

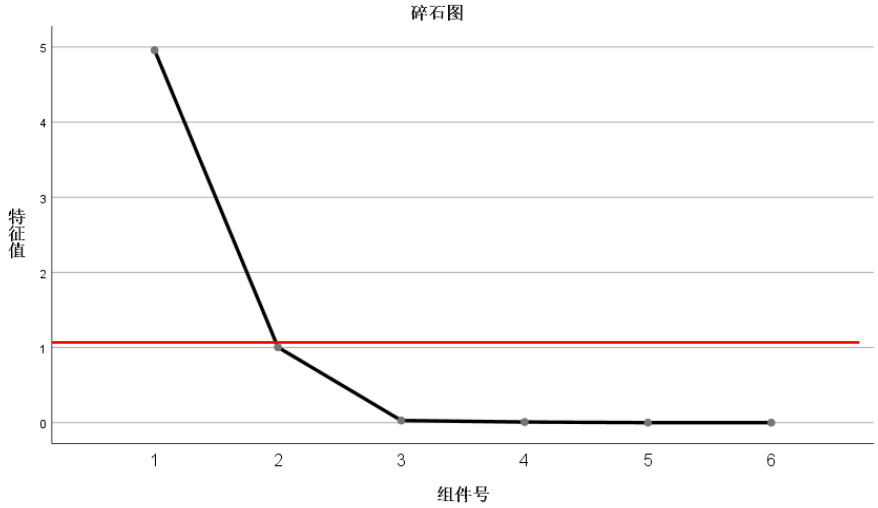

图 4.9 因子分析碎石图(2016 年)

根据上述的分析结果，我们可以建立 2016 年关于上述 6 个评价指标的恐怖袭击事件风险评估模型，最终可以得到恐怖事件的风险指数。根据我们对各指标的量化原则可以得到，风险指数越大，该地区发生恐怖袭击事件的风险越大；相反，风险指数越小则表示该地区发生恐怖袭击事件的风险越小。2016 的恐怖袭击事件风险评估模型如下所示：

$$Y = 0.191 * S_1 + 0.192 * S_2 + 0.188 * S_3 + 0.187 * S_4 + 0.190 * S_5 + 0.052 * S_6$$

3、2017 年恐怖袭击事件分析

我们首先分析 2017 年的数据，对其进行因子分析。因子分析结果如表 4.8 和图 4.10 所示。

表 4.8 因子分析结果(2017 年)

	初始特征值			提取载荷平方和		
成分	总计	方差百分比	累积%	总计	方差百分比	累积%
1	4.981	83.015	83.015	4.981	83.015	83.015
2	0.858	14.296	97.312			
3	0.122	2.029	99.341			
4	0.022	0.373	99.714			
5	0.017	0.283	99.997			
6	0	0.003	1			

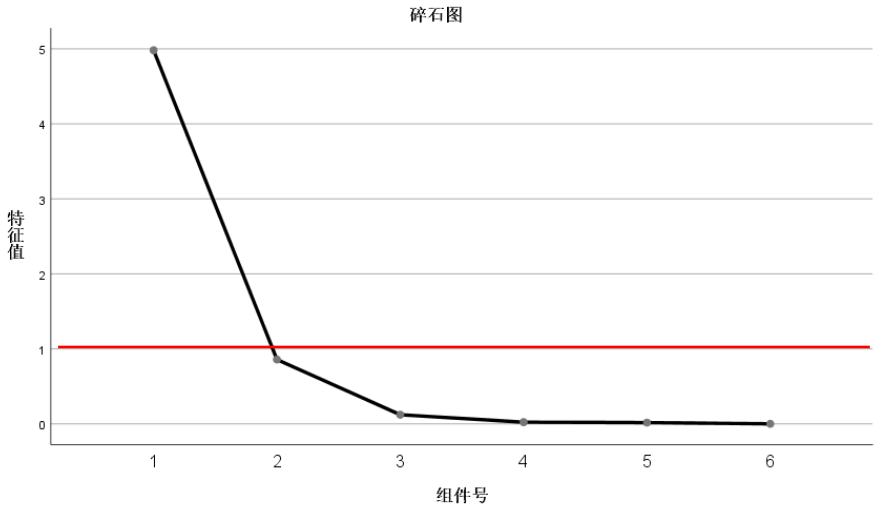

图 4.10 因子分析碎石图(2017 年)

根据上述的分析结果，我们可以建立 2017 年关于上述 6 个评价指标的恐怖袭击事件风险评估模型，最终可以得到恐怖事件的风险指数。根据我们对各指标的量化原则可以得到，风险指数越大，该地区发生恐怖袭击事件的风险越大；相反，风险指数越小则表示该地区发生恐怖袭击事件的风险越小。2017 的恐怖袭击事件风险评估模型如下所示：

$$Y = 0.187 * S_1 + 0.186 * S_2 + 0.184 * S_3 + 0.185 * S_4 + 0.179 * S_5 + 0.079 * S_6$$

4.5 恐怖袭击事件综合风险分析

根据 4.4 节建立的恐怖袭击事件风险评估模型分别计算 2015 年、2016 年和 2017 年各个地区的风险评分，并对其进行排序，如表 4.9 所示。

表 4.9 不同地区的风险评分及排名(2015~2017)

地区	2015		2016		2017	
	风险评分	排名	风险评分	排名	风险评分	排名
NA	46.788	9	71.204	8	230.365	5
CAC	0.293	12	3.098	12	2.765	12
SA	117.438	7	108.598	6	115.648	7
EA	47.582	8	16.895	9	19.285	9
SEA	823.569	4	694.633	4	724.722	4
SAS	5088.108	2	4441.198	2	4273.064	2
CA	8.04	10	15.099	10	5.057	11
WE	228.676	6	290.551	5	198.176	6
EE	649.388	5	86.829	7	93.372	8

MENA	9917.179	1	10065.56	1	5178.666	1
SSA	3878.099	3	2874.309	3	2854.001	3
AO	5.446	11	3.966	11	9.306	10

从表 4.9 中可以看出，中东、南亚以及非洲始终处于恐怖袭击事件风险评估的前三名，且风险评分相较于 2 地区差别明显，因此我们可以看出，目前全球绝大部分的恐怖袭击事件仍发生在这三个地区。从中我们也可以看出虽然这三个地区处于国际恐怖袭击时间发生地的前三名，但它们的风险评估评分逐年下降，说明这三个地区的恐怖袭击活动正逐渐受到遏制，反恐态势有所好转，但仍不可掉以轻心。另一个显著的变化就是北美地区，2015 年北美地区风险评估排名第 9，但到了 2017 年，它的风险评估排名已经达到了第 5 名，且风险评分逐年增加。

从空间角度看，中东、南亚以及非洲为恐怖袭击的高发风险地区；东南亚、东欧和西欧为中等风险地区；其余地区为低风险地区。但从 2015 年至 2017 年的数据可以看出，高风险地区的恐怖袭击事件发生的风险正在逐年降低，但北美地区的风险正在增加，恐怖袭击风险从南亚和中东转移到了北美地区。

4.6 未来反恐态势分析

表 4.10 恐怖袭击统计信息(2015~2017)

年度	2015	2016	2017
年度总次数	14963	13587	10900
年度成功次数	12674	10975	8652
成功百分比	0.84702266	0.807757415	0.793761468
总死亡人数	38851	34871	26445
受伤人数	44037	40001	24927
财产损失(美元)	3352187.45	211894186.7	13154956.03
绑架人数	29	7	10
绑架时间	244	200.25	167
索要赎金	193008519	35929524.42	20248525.61
自杀式袭击	922	985	844

从表 4.10 可以看出，在 2015 年至 2017 年这三年间，随着年份的增加，恐怖袭击的次数由 14963 次降低到了 10900 次，成功的袭击次数从 12674 次降低到了 8652 次，成功率由 84.7%降低到了 79.3%。同时受伤人数和死亡人数也随着年份的增加有了大幅的减少。从中我们可以得出以下结论：全球的反恐态势显变好的趋势，无论是从恐怖袭击的发生次数，还是恐怖袭击造成的伤亡人数，均有明显的降低，因此预计 2018 年的国际反恐态势，但仍不可掉以轻心。

根据我们本节所建立的恐怖袭击事件风险评估模型得到了不同地区在 2015 年至 2017 年的风险评估得分及其相应的排名。从表 4.9 中可以看出，中东、南

亚以及非洲然是目前国际上恐怖袭击事件最多发的区域，约占全球所有恐怖袭击事件的 85%，但它们的风险评估得分逐年减少，可见在这些恐怖袭击事件的局部高发区域，风险有所降低。另外，在北美区域，虽然它们本土的恐怖袭击事件的风险评估得分相较于中东、南亚等区域具有较大的差别，但它们的风险评估得分逐年增加，并且风险排名也从 2015 年的第 9 名上升到了 2017 年的第 5 名。因此我们可以预计在 2018 年，中东、南亚和非洲的恐怖袭击事件风险会继续向北美地区转移，其他地区保持平稳。但中东、南亚等区域仍是目前全球恐怖袭击事件最频发的区域。

4.7 针对反恐斗争的建议

1、根据分析可以得知，北美地区的风险逐年增加。由于美国地区没有禁枪制度，导致有大量的枪支分布在民间，这也是北美地区近年恐怖袭击事件比例有所增加的原因。因此，应该完善枪支、涉爆物品的管理制度，加强对涉枪犯罪的打击力度。

2、中东、南亚以及非洲地区是目前国际恐怖袭击事件最频发的三个区域，尽管近几年的比例有所下降，但仍占据了绝大部分。因此针对这些重点区域，要加强恐怖组织的情报监测，并且及时预警。保持严打高压态势，定期展开恐怖组织的清剿活动。

第五章 任务四解答

在本章中，主要通过数据分析了中国恐怖主义事件的特征，通过相关性分析分析了中国恐怖主义事件随年变化与国际恐怖主义事件发展变化之间的相关性，从主要地区，主要恐怖组织，攻击方式，武器类型等方面介绍了中国恐怖主义时间的特征，并且针对这些特征给出了对应的防控措施。通过数据分析给出了通过恐怖组织之间活动的相互联系，预测恐怖活动的方法模型[9][10]。

5.1 中国恐怖事件分析

5.1.1 中国恐怖袭击与全球恐怖袭击数量之间的关系

近年来，全球恐怖袭击频发，中国也发生了一些恐怖袭击事件，根据 GTD 数据库中的记录，1998-2017 年之间的 20 年之间，全球发生了 114813 起恐怖袭击事件，中国也发生了共 141 起恐怖袭击事件，对 10 年间的恐怖袭击事件进行统计，统计结果如下：

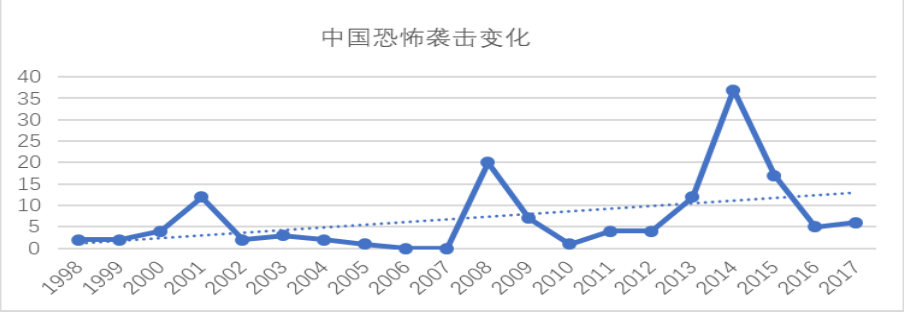

图 4.1.1.1 中国历年恐怖袭击事件变化

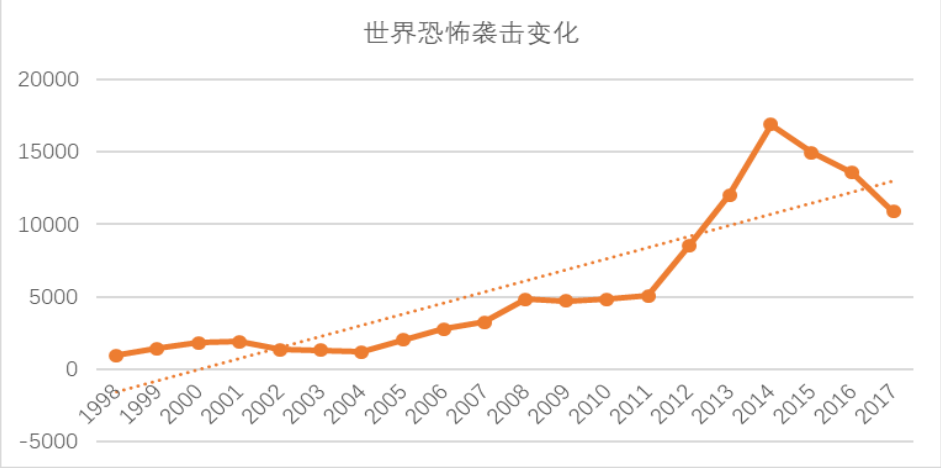

图 4.1.1.2 世界历年恐怖袭击事件变化

由上图可以看出，中国和世界的恐怖袭击事件都在逐渐增加，在 2014 年达到最高值，且在 2001 年，2008 年和 2014 年，恐怖袭击事件的数量有一个局部

的最高点，在这三年后面的一定时间内袭击事件的数量出现了一定程度上的减少，在整体上中国和世界的恐怖袭击事件的数量达到趋势的统一。

对两者的数量进行定量分析，在 SPSS 中进行相关性分析，分析结果如下表所示，两者的相关性达到 0.663，相关性比较显著。

表 5.1.1.1 相关性分析

		中国恐怖袭击次数	世界恐怖袭击次数
中国恐怖袭击次数	皮尔逊相关性	1	.663**
	Sig.（双尾）		.001
	个案数	20	20
世界恐怖袭击次数	皮尔逊相关性	.663**	1
	Sig.（双尾）	.001	
	个案数	20	20

**. 在 0.01 级别（双尾），相关性显著。

5.1.2 中国恐怖袭击地区分析

1998 年以来，中国国内发生的 141 起恐怖袭击事件，涉及了中国的 23 个省市，其中在新疆发生 69 起恐怖袭击事件，广东发生了 11 起恐怖袭击事件，北京发生了 9 起恐怖袭击事件，云南 7 起。造成了 893 人死亡，1296 人受伤，对社会造成了十分恶劣的影响。统计结果如图所示：

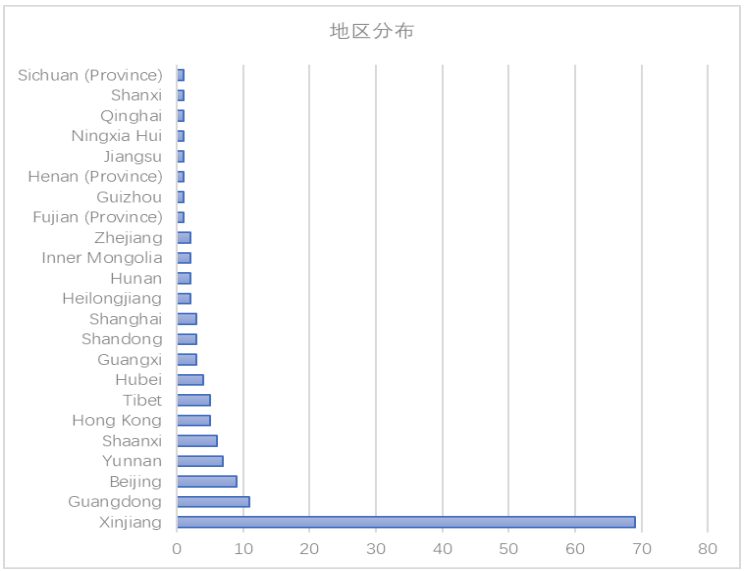

图 5.1.2.1 地区恐怖袭击数量统计

其中恐怖事件的频发地点是新疆，广东，北京，云南，陕西，香港和西藏，

这些省份都是靠近中亚和南亚地区的边界省份，南亚作为恐怖事件的多发地带，中国国内的恐怖事件频发一定程度上受到了国际的影响。新疆的袭击次数统计如下表所示：

表 5.1.2.1 新疆恐怖袭击次数统计

年份	袭击次数	年份	袭击次数
1998	1	2012	2
2001	1	2013	10
2005	1	2014	29
2008	4	2015	11
2009	3	2016	1
2010	1	2017	1
2011	3		

5.1.3 恐怖分子分析

由图 5.1.2.1 可以看出，新疆的恐怖袭击数量占据主要的部分，在新疆发生的这 69 起恐怖袭击事件中，其中有 35 起的恐怖袭击的制造者未知，剩下的 34 起恐怖袭击事件中，其中 28 起由维吾尔族分裂主义者制造，3 起由突厥斯坦伊斯兰党制造，3 起由东突厥斯坦伊斯兰运动（ETIM）组织制造。广东省的 11 起恐怖袭击事件其中两起由维吾尔族分裂者制造，在云南发生的 7 起恐怖袭击事件其中 2 起由突厥斯坦伊斯兰党制造，1 起由维吾尔族分裂主义者制造，2 起由克钦独立军和全国民主联盟军制造。通过对恐怖袭击制造者的统计，统计结果如下表所示：

表 5.1.3.1 恐怖组织统计

恐怖组织	恐怖事件	死亡人数	受伤人数
Uighur Separatists	34	359	441
Turkestan Islamic Party	6	107	146
Eastern Turkistan Islamic Movement (ETIM)	4	25	32
Kachin Independence Army (KIA)	2	0	1
Tibetan separatists	1	1	0
Unknown	94	401	676

由表可以看出，维吾尔族分裂者制造的 34 起恐怖袭击共造成 359 人死亡，441 人受伤，由突厥斯坦伊斯兰党制造的 6 起恐怖袭击共造成 107 人死亡，146 人受伤，由此可以看出，恐怖主义组织的残暴之处。

5.1.4 恐怖袭击方式和对象分析

对 GTD 中提供的恐怖袭击的方式进行统计，统计结果如下表所示，通过分析表中数据可以看出，恐怖组织的主要攻击方式是爆炸和武装突击，这两种方式具有攻击性强和伤害比较大的特点。由于采用爆炸的攻击方式更能造成更大的伤害，容易对社会和民众产生更大的震慑力，因此，这两种攻击方式是主要的攻击方式。

表 5.1.4.1 攻击方式统计

attacking type	次数
Bombing/Explosion	77
Armed Assault	43
Unarmed Assault	6
Facility/Infrastructure Attack	5
Hijacking	5
Assassination	2
Hostage Taking (Kidnapping)	1
Unknown	2

对恐怖分子的武器类型进行分析，主要有炸药，持械乱斗，防火，化学武器以及车辆撞击，这些武器类型更容易在大量的民众中造成震慑，因此恐怖分子多采用这些武器，在 20 年间出现了 3 次化学武器的使用，统计结果如下表所示：

表 5.1.4.2 武器类型统计

weapon type	次数
Explosives	85
Melee	31
Incendiary	17
Chemical	3
Unknown	2
Vehicle (not to include vehicle-borne explosives, i.e., car or truck bombs)	2
Firearms	1

恐怖分子的攻击对象一般以警察，私人民众，商业场所，公共交通和政府机构为主，以便于达到震慑和扩大影响的目的，而且一般的普通民众没有足够的自保能力，面对恐怖袭击更加容易被控制，因此民众，商业场所和公共交通成为恐怖袭击的主要对象。恐怖袭击对象统计如下表所示：

表 5.1.4.3 攻击对象统计

targtype	次数
Police	36
Private Citizens & Property	36
Business	23
Transportation	17
Government (General)	11
Airports & Aircraft	5
Educational Institution	4
Military	2

Journalists & Media	1
Religious Figures/Institutions	1
Tourists	1
Unknown	4

5.1.5 我国恐怖袭击的防控措施

1、针对重点区域和重点对象进行具体的防控

在前面的分析中，恐怖主义主要发生在边境省份，攻击的对象为公众场合，交通工具，民众，以及政府和警察机构，针对这些攻击对象，需要采用的措施有，加大公共场合的安全检查，拒绝危险物品的进入，并且加大对危险的可疑分子的检查，在各地的汽车，火车，飞机等交通工具的枢纽点进行乘客的安全检查。加强对相关政府机构的装备，为及时发现和制止恐怖袭击做准备。

2、针对武器源头进行防控

相比其他国家而言，中国的恐怖袭击事件采用枪支等武器的事件比较少，说明中国的禁止枪支的规定获得了较好的成果。禁止枪支的规定应该长久的执行下去。爆炸物的使用是恐怖袭击事件的主要方式之一，为了减少爆炸物的后果，应该加大对爆炸物危害的宣传，从源头控制爆炸物的产生。

3、针对边界地区与国际反恐组织进行合作

边界地区靠近易发生恐怖空袭的南亚等地方，为了防止国内受到邻国恐怖组织的渗透和影响，应该与国际反恐组织进行合作，消灭恐怖组织，实现对恐怖事件的预防。

5.2 根据历史数据进行未来恐怖行动的预测

5.2.1 犯罪集团分析

犯罪集团的犯罪行为具有一定的组织性，大型的犯罪集团犯罪会吸引小型的犯罪集团跟随，历史数据中往往包含着一些比较典型的犯罪规律。

5.2.1.1 对第一犯罪集团和第二犯罪集团进行相关性分析

选择第一犯罪集团和第二犯罪集团进行相关性分析，相关性分析如下表所示：

表 4.2.1.1 相关性分析

		Gname1	Gname2
Gname1	皮尔逊相关性	1	.604**
	Sig.（双尾）		.000
	个案数	1376	1376
Gname2	皮尔逊相关性	.604**	1
	Sig.（双尾）	.000	
	个案数	1376	1376

**. 在 0.01 级别（双尾），相关性显著。

表中数据显示，两者的相关性达到 0.604，具有较高的相关性。因此考虑寻找这两者之间的关系，进行建模。

5.2.1.2 犯罪集团特征分析

取犯罪集团的特征进行分析，不同地区、规模的组织往往采用不同的武器，着重攻击不同的场合。因此在预谋犯罪时，也会跟这些信息相关，包括武器的购入，路线的查探，声明方式等相关，可以通过这些途径分析不同组织的犯罪概率。

5.2.2 模型建立

5.2.2.1 建模分析

假设第一犯罪组织为主要的犯罪组织，第二第三犯罪组织为从犯，由于作为第一犯罪组织的对象具有更大的号召力，因此会有比较多的小型组织作为跟随。考虑将第一犯罪组织定义为 $x_1, x_2 \cdots x_n$，与这些第一犯罪组织相关的第二犯罪组织定义为 $y_1, y_2 \cdots y_m$，两者之间的关系图如下：

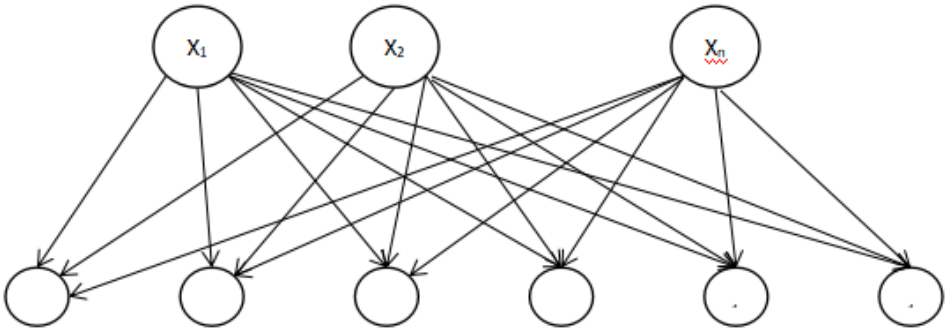

图 4.2.2.1 拓扑关系

其中 $x_i(i=1...n)$ 与 $y_j(j=1...m)$ 之间相互拓扑，表示第一犯罪集团和第二犯罪集团之间的相互联系。

首先根据历史攻击信息，可以求出犯罪组织之间的联系关系，联系关系可以通过计算第二组织与第一组织之间的攻击次数相关性来得到相互联系的权值，假设 $y_j(j=1...m)$ 与 $x_i(i=1...n)$ 之间的正向权重值为 w_{ij}。

$$w_{ij} = \frac{z_{ji}}{\sum_{i=1}^{n} z_{ji}}$$

其中，$z_{ij}(j=1...m)$ 表示 $y_j(j=1...m)$ 跟随 $x_i(i=1...n)$ 进行犯罪活动的次数，依照以上的公式计算出第一犯罪集团与第二犯罪集团之间反向权值，即第一犯罪组

织在第二犯罪组织的占有权重，即第二犯罪组织对于第一犯罪组织的依附程度，w_{ij} 越大，表示这个第二集团对于该第一集团的依附程度越高。

然后进行反向推导，假设 $x_i(i=1...n)$ 与 $y_j(j=1...m)$ 之间的正向权重值为 v_{ij}。

$$v_{ij}=\frac{w_{ji}}{\sum_{j=1}^{m} w_{ji}}$$

正向权重值通过对同一个第一犯罪组织对多个第二犯罪组织的反向权值的占有比例来获得。v_{ij} 越大，表示第二集团对该第一集团的依附程度越高。

假设某第一犯罪集团 $x_p(1<=p<=n)$ 正在准备犯罪活动，对 m 个依附于它的第二集团进行行为分析，包括武器收购，网络或者人际联络，分别获取 m 个第二集团的犯罪概率 $O_j(j=1...m)$，然后通过反向权值来计算第一集团的犯罪概率

$$M_i=\sum_{j=1}^{m} O_j V_{ij}$$

通过第一集团的犯罪概率获得最高可能犯罪的组织，然后进行有效的防范。

5.2.2.2 算法升级

通过以上的权值计算方法会有一定的误差性，为了提高预测的概率，更精确的了解两种集团之间的联系，考虑采用神经网络的方法来求取双向的联系权值。

将第一犯罪组织作为神经网络的输入层，第二犯罪组织作为神经网络的输出层，隐含层采用一定的变量，首先通过正向训练求取正向权值，然后反向求取，直到训练误差降低到满足实验的条件。

然后通过上面的建模方法，通过分析第二组织的活动状况，来监测第一集团的犯罪动机。

5.2.2.3 可行性分析

1) 第一集团和第二集团具有一定的相关性。
2) 可以通过神经网络降低双向权值的误差。

参考文献

[1] 王雷, 王欣, 赵秋红. 基于和声搜索算法优化支持向量机的突发暴恐事件分级研究[J]. 管理评论, 2016, 28(8):125-132.
[2] 李德仁. 空间数据挖掘理论与应用[M]. 科学出版社, 2013.
[3] 王晓东, 许占文. 高效关联规则数据挖掘算法研究[J]. 沈阳工业大学学报, 2002, 24(4):329-333.
[4] 陆瑶. 大数据背景下的小米手机营销策略研究[D]. 江苏大学, 2016.
[5] 李旭. 五种决策树算法的比较研究[D]. 大连理工大学, 2011.
[6] 郭开明. 基于 GTD 数据库的国内恐怖活动现状及防控对策研究[J]. 江西警察学院学报, 2018(2).
[7] 王锂达, 张闯. 挖掘恐怖组织行为模式:基于 FP-Growth 的滑动时间窗口衰减算法研究[J]. 2016.
[8] 赵法栋, 庄弘炜, 金振兴. 基于 MLE 的恐怖组织袭击行为模式实证研究[J]. 复杂系统与复杂性科学, 2014, 11(4):19-22.
[9] 王伟. 基于背景知识的恐怖行为预测算法研究[D]. 江苏大学, 2012.
[10] 战兵, 韩锐. 基于隐马尔可夫的恐怖事件预测模型[J]. 解放军理工大学学报(自然科学版), 2015(4):386-393.

附录

附录一：决策树准确率统计程序

```
%%
% num1 = xlsread('C:\Users\xiaofan\Desktop\Sourced_Data\Test_data2017.xlsx');
% num2 = xlsread('C:\Users\xiaofan\Desktop\Sourced_Data\traindata_revised.xlsx');
sum_test = size(num1,1);
sum_train =   size(num2,1);
train_index =1;
test_index = 1;
accuracy = 0;
error = 0;
accuracy_rate = 0;
have_same_index = 0;

for test_index=1:1:sum_test
    for train_index = 1:1:sum_train
       if strcmp(num1(test_index,1),num2(train_index,1))
          if num1(test_index,7) == num2(train_index,6)
             accuracy = accuracy+1;
             continue;
          else   if num1(test_index,7) ~= num2(train_index,6)
             error = error+1;
             continue;
             end
          end
       end
    end
    fprintf ('%d',test_index);
end
accuracy_rate = accuracy/sum_test;
```

附录二：量化分级模型危害性得分程序

```
% clear

x1_1 = data1(:,3);
x1_2 = data1(:,4);
x1_3 = data1(:,5);
x1_4 = data1(:,6);
x1_5 = data1(:,7);
x1_6 = data1(:,8);
x1_7 = data1(:,9);
```

```
x1_8 = data1(:,10);
x1_9 = data1(:,11);
x1_10 = data1(:,12);
x1_11 = data1(:,13);

k1=length(data1);

for i=1:k1
      y1(i) =
0.0512*x1_1(i)+0.003*x1_4(i)+0.061*x1_5(i)+0.179*x1_6(i)+0.177*x1_8(i)+0.242*
x1_9(i)+0.239*x1_10(i)+0.048*x1_11(i);
end

x2_1 = data2(:,3);
x2_2 = data2(:,4);
x2_3 = data2(:,5);
x2_4 = data2(:,6);
x2_5 = data2(:,7);
x2_6 = data2(:,8);
x2_7 = data2(:,9);
x2_8 = data2(:,10);
x2_9 = data2(:,11);
x2_10 = data2(:,12);
x2_11 = data2(:,13);
k2=length(data2);

for i=1:k2
      y2(i) =
-0.024*x2_2(i)+0.111*x2_3(i)-0.081*x2_4(i)+0.152*x2_5(i)+0.264*x2_6(i)+0.084*
x2_7(i)+0.283*x2_8(i)+0.051*x2_9(i)+0.115*x2_10(i);
end

x3_1 = data3(:,3);
x3_2 = data3(:,4);
x3_3 = data3(:,5);
x3_4 = data3(:,6);
x3_5 = data3(:,7);
x3_6 = data3(:,8);
x3_7 = data3(:,9);
x3_8 = data3(:,10);
x3_9 = data3(:,11);
x3_10 = data3(:,12);
x3_11 = data3(:,13);
k3=length(data3);
```

```
for i=1:k3
      y3(i) =
0.14*x3_1(i)-0.011*x3_2(i)+0.121*x3_3(i)-0.156*x3_4(i)+0.141*x3_5(i)+0.167*x3_
6(i)+0.193*x3_8(i)+0.108*x3_9(i)+0.12*x3_10(i)+0.177*x3_11(i);
end

x4_1 = data4(:,3);
x4_2 = data4(:,4);
x4_3 = data4(:,5);
x4_4 = data4(:,6);
x4_5 = data4(:,7);
x4_6 = data4(:,8);
x4_7 = data4(:,9);
x4_8 = data4(:,10);
x4_9 = data4(:,11);
x4_10 = data4(:,12);
x4_11 = data4(:,13);

k4=length(data4);

for i=1:k4
      y4(i) =
-0.093*x4_1(i)+0.099*x4_2(i)+0.138*x4_4(i)+0.294*x4_6(i)+0.293*x4_8(i)+0.166*
x4_9(i)+0.142*x4_10(i);
end


x5_1 = data5(:,3);
x5_2 = data5(:,4);
x5_3 = data5(:,5);
x5_4 = data5(:,6);
x5_5 = data5(:,7);
x5_6 = data5(:,8);
x5_7 = data5(:,9);
x5_8 = data5(:,10);
x5_9 = data5(:,11);
x5_10 = data5(:,12);
x5_11 = data5(:,13);

k5=length(data5);

for i=1:k5
      y5(i) =
```

```
0.127*x5_1(i)+0.091*x5_2(i)+0.059*x5_3(i)-0.11*x5_4(i)+0.11*x5_5(i)+0.222*x5_
6(i)+0.227*x5_8(i)+0.123*x5_9(i)+0.128*x5_10(i);
end

[a,k]    = sort(y1);
[b,q] = sort(y2);
[c,w] = sort(y3);
[d,o] = sort(y4);
[e,g] = sort(y5);
```