“华为杯”第十五届中国研究生数学建模竞赛

学　　校	华东理工大学　华东师范大学
参赛队号	**18102510035**
队员姓名	**1.** 沈皓杰
	2. 李佳倩
	3. 张文清

“华为杯”第十五届中国研究生
数学建模竞赛

题　　目　全国研究生数学建模竞赛论文标题

摘要

还没写

目录

1. 问题重述

1.1 问题背景

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、不符合国际道义的攻击行为。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 需要解决的问题

现有数据如下：附件 1 选取了某组织搜集整理的全球恐怖主义数据库 (GTD) 中 1998-2017 年世界上发生的恐怖袭击事件的记录；附件 2 是有关变量的说明，节译自数据库说明文档。附件 2 文档较长，附件 3 提供了一个内容摘要。以此 3 份附件为数据基础，本文依次解决如下问题：

问题 1：依据附件 1 以及其它有关信息，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，将附件 1 给出的事件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。

表 **1**　典型事件危害级别

事件编号	危害级别
200108110012	
200511180002	
200901170021	
201402110015	
201405010071	
201411070002	
201412160041	
201508010015	
201705080012	

问题 2：依据事件特征发现恐怖袭击事件制造者。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，运用数学建模方法寻找并案调查可能性，即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（表中样例的意思是：对事件编号为 XX 的事件，3 号的嫌疑最大，其次是 4 号，最后是 5 号），如果认为某嫌疑人关系不大，也可以保留空格。

表 **2**　恐怖份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					
201702210037					
201703120023					
201705050009					
201705050010					
201707010028					
201707020006					
201708110018					
201711010006					
201712010003					

问题 3：依据附件 1 并结合因特网上的有关信息，建立适当的数学模型，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，用图表给出你们的研究结果，提出你们对反恐斗争的见解和建议。

问题 4 给出模型和方法，通过数学建模进一步挖掘附件 1 数据的作用。

2. 模型的假设

假设 1 高维数据 (样本) 位于一个维数比数据空间维数小得多的流形上；

假设 2 高维数据空间到低维数据空间上的映射是线性映射；

假设 3 不考虑少数样本的特有维度维度对恐怖袭击事件危害程度分级的影响；

3. 符号说明

符号	意义
D	木条宽度（cm）
L	木板长度（cm）
W	木板宽度（cm）
N	第 n 根木条
T	木条根数
H	桌子高度（cm）
R	桌子半径（cm）
R	桌子直径（cm）

4. 问题一

4.1 问题分析

问题一要求基于附件 1 给出的恐怖袭击事件记录数据，通过量化分析，将事件按危害程度分级。已知分类信息的历史数据并不完全，因此这是一个典型的数据挖掘聚类问题。

由于附件提供的数据维数很高，数据结构比较复杂，难以将其直观地与恐怖袭击事件的危害程度联系起来。然而对上述数据进行分析时，并非所有属性对随后进行的数据处理都是重要的。高维数据中的信息往往包含在一个或几个低维结构中，可以用少量的简单变量来支配。

因此首先要对数据进行预处理，流程包括数据清理 (包括对不完整数据的处理)、数据噪声消除、数据变换 (标准化) 等。数据经过预处理后，仍保留了其原始的特征和规律，通过主成分分析 (Principle Component Analysis,PCA) 约减维数可以更好地进行分析。完全定性的聚类方法太过主观；定量聚类方法虽然客观，但由于舍弃了无法量化的信息，会导致结果不符合实际。本文定性与定量相结合，通过基于贝叶斯聚类算法 (Bayesian Hierarchical Clustering) 的多模块贝叶斯网络方法来进行聚类，最后达到分级的目的。

4.2 模型建立

4.2.1 数据预处理

首先对全球恐怖主义数据库 (GTD) 中 1998-2017 年世界上发生的恐怖袭击事件进行整理，整理步骤如下：

第 1 步 提取确定是恐怖袭击事件的样本；

第 2 步 根据事件摘要来选取与恐怖袭击事件危害程度相关的数据；

第 3 步 将含有缺失数据的样本剔除；

第 4 步 因原数据集中恐怖袭击所致死亡人数和伤亡人数中包含了恐怖分子的死亡和伤亡人数，而本文评估的死伤人数不包含恐怖分子，所以用总的死伤人数分别减去恐怖分子死伤人数得到修正的死亡和伤亡人数;

第 5 步 因第 1 步和第 3 步的操作，使得数据集中某些属性的取值不连续，找到这些属性，并将它们的取值连续化。

经过以上整理，得到具有 95711 条样本量的完整数据集。

4.2.2 数据降维

降维是构造降维映射，获得高维数据低维表示的方法，是对大量高维无序且没有明显空间特征数据的处理，以发现隐藏在高维数据中有意义的低维结构。降

维问题假设高维数据 (样本) 位于一个维数比数据空间维数小得多的流形上，降维的目的就是获得这一流形的低维坐标表示。设 $X=(x_1,x_2,...x_D)^T$ 是高维空间 R^D 中的向量，通过降维

$$F(x)=\begin{pmatrix}F_1(X)\\F_2(X)\\\vdots\\F_d(X)\end{pmatrix}=\begin{pmatrix}F_1(x_1,x_2,\cdots,x_D)\\F_2(x_1,x_2,\cdots,x_D)\\\vdots\\F_d(x_1,x_2,\cdots,x_D)\end{pmatrix}. \tag{1}$$

得到低维空间 R^d 中的向量 $Y=(y_1,y_2,\cdots,y_d)^T$。本文假设高维数据到低维空间上的映射是线性映射。因此对于问题一，本文采用主成分分析 (Principle Component Analysis,PCA) 来进行降维。PCA 将数据的主成分 (包含信息量大的维度) 保留下来，忽略掉对数据描述不重要的成分。即将主成分维度组成的向量空间作为低维空间，将高维数据投影到这个空间上就完成了降维的工作。

设样本数目为 n, 每个样本观测 p 项指标 (即维数为 p)x_1,x_2,$\cdots$,x_p, 则根据附件 1 资料提供的数据标准化处理得到的原始数据矩阵为：

$$X_{np}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&\vdots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{np}\end{pmatrix}\equiv\begin{pmatrix}X_1&X_2&\cdots&X_p\end{pmatrix}. \tag{2}$$

用数据矩阵 X_{np} 的 p 个向量 X_1,X_2,X_p 作线性组合为：

$$F_i=a_{i1}X_1+a_{i2}X_2+\cdots+a_{ip}X_p(i=1,2,\cdots,p). \tag{3}$$

其中，系数 a_{ij} 由下列原则来确定：

1.F_i 与 F_j 互不相关。

2.F_1 是 x_1,x_2,$\cdots$,x_p 的一切线性组合且满足 (3) 中方差最大的;F_2 是与 F_1 不相关的 x_1,x_2,$\cdots$,x_p 的所有线性组合中方差最大者;F_p 是与 F_1，F_2,$\cdots$,F_{p-1} 都不相关的 x_1,x_2,$\cdots$,x_p 的一切线性组合中方差最大的。

根据定理，由 X 求得的协方差矩阵就是相关系数矩阵

$$A=\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1p}\\a_{21}&a_{22}&\cdots&a_{2p}\\\vdots&\vdots&\vdots&\vdots\\a_{p1}&a_{p2}&\cdots&a_{pp}\end{pmatrix}. \tag{4}$$

求 A 的特征值和特征向量可通过正交变换 T 使

$$T^{-1}AT = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}. \tag{5}$$

按照累积方差贡献率 $\frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{p}r_{jj}} = \frac{\sum_{j=1}^{k}\lambda_j}{p} > 85\%$ 从而建立前 k 个主成分：

$$F_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p (i = 1, 2, \cdots, k). \tag{6}$$

4.2.3 数据聚类

本文将用于恐怖袭击危害程度分级的贝叶斯网络分为 4 个模块，分别是恐怖袭击危害程度的分级贝叶斯网络模块以及 3 个后果子模块：财产损失模块、人员伤亡模块和不良社会影响模块。图1是恐怖袭击危害程度的分级贝叶斯网络模块，其中的置换节点可以被相应的贝叶斯网络模块所置换。HL(harm level) 表示危害程度，PD(property damage) 和 CeP(casualties except perpetrators) 是恐怖主义活动造成财产损失和人员伤亡情况, 并且人员伤亡不包括恐怖份子。ASI(adverse social impacts) 是恐怖主义行为造成的社会不良影响程度。

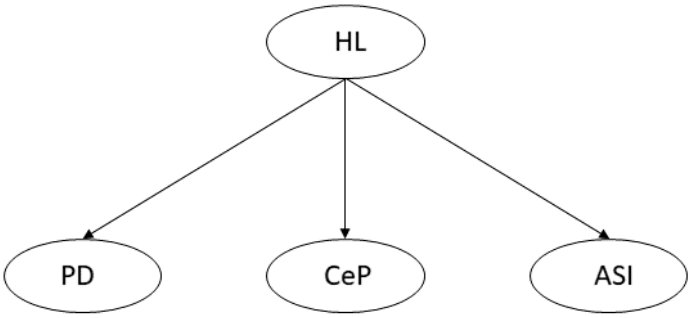

图 **1**　分级贝叶斯网络模块

由于财产损失模块和人员伤亡模块的样本数据具有特征 (feature) 和标签 (label)。因此对于这 2 个模块，我们采用监督学习的方法来构建，本质上就是找到特征和标签间的映射。对于不良社会影响模块，根据专家知识来构建较为合理。

本文先对人员伤亡贝叶斯网络模块的构建进行分析，财产损失的构建方法类似。人员伤亡贝叶斯网络模块中的节点，表示对人员伤亡评估有意义的情报信

息或者经过情报机构的专业人员通过数据融合获得的信息。表**??**表示对于人员伤亡模块，有人员伤亡 1 个类节点和 15 个属性节点。

对于本文的人员伤亡模块，其节点变量较少，所以文本用打分搜索方法中的 K2 算法来构建贝叶斯网络。这个算法需要指定结点的先验顺序。节点顺序主要包括领域知识或节点偏序指定的约束。为了确定唯一的初始节点序，我们首先基于定向后的最大权重跨度树对节点块排序，结点块的排序是唯一的，然后再采用局部完全有向无环图法对块内节点排序，最终得到所有节点的排序结构如图。同理可得财产损失模块的贝叶斯网络结构如图所示。

其次对不良社会影响模块进行分析。根据专家知识可知，恐怖袭击事件造成的不良社会影响跟恐怖袭击事件造成的人员伤亡、财产损失，恐怖主义活动发生前是否威胁、恐吓或向大众传播某种恐怖消息，以及民众或反恐决策人员对恐怖袭击的知情程度密切相关。由此得到图2所示的不良社会影响模块的贝叶斯网络结构。TA(terrorist atmosphere) 是恐怖份子是否在恐怖活动前对大众造成恐吓或威胁；DK(degree of knowledge) 是民众对恐怖袭击的知情程度。PD 和 TA 的状态合集如**??**所示。根据专家知识，ASI,CeP 和 DK 变量的状态合集如表：

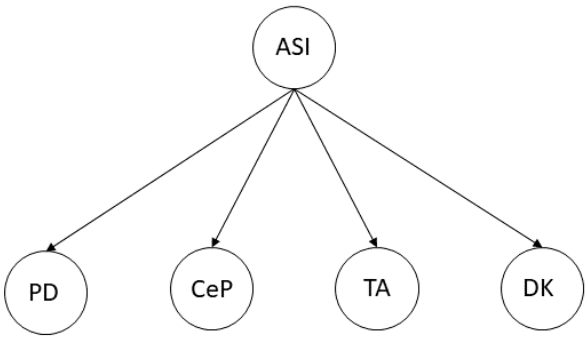

图 **2**　不良社会影响模块的贝叶斯网络结构

设贝叶斯网络中某一节点 R, 它有 n 个取值，每个取值的先验概率为 $p(r_i)$, 且 $\sum_{i=1}^{n} p(r_i) = 1$, 节点 R 有一取 m 个值的子节点 S,$p(S|R)$ 的条件概率矩阵为

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{pmatrix}. \tag{7}$$

财产损失模块和人员伤亡模块贝叶斯网络，由于存在观测数据，可以从数据中学习参数的概率分布表，但没有与分级贝叶斯网络和不良社会影响模块贝叶

斯网络相关的观测数据，所以这 2 个贝叶斯网络模块的概率分布表采用主观赋权法来确定。考虑多个专家给出 $p(S|R)$ 的条件概率矩阵为

$$P^k = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{pmatrix}. \tag{8}$$

其中 $k = 1, 2, \cdots, q$ 表示专家的个数，设每个专家的权重为 $H = (h_1, h_2, \cdots, h_q)^T$，其中 $\sum_{k=1}^{q} h_k = 1$. 为了确定各条件概率值，建立优化模型：

$$\begin{aligned} \min &= \sum_{i=1}^{n} p(a_i) \sum_{k=1}^{q} \sum_{j=1}^{m} h_k (p_{ij} - p_{ij}^k)^2 \\ \text{s.t.} &\sum_{j=1}^{m} p_{ij} = 1 (i = 1, 2, \cdots, n), \\ p_{ij} &\geq 0 (i = 1, 2, \cdots, n; j = i = 1, 2, \cdots, m) \end{aligned} \tag{9}$$

该优化模型的目标是使条件概率矩阵与专家给出的条件概率之间的总偏差平方和最小。由 $\frac{\partial L}{\partial p_{ij}}$ 得条件概率的主观概率为

$$p_{ij} = \sum_{k=1}^{q} h_k p_{ij}^k (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m). \tag{10}$$

4.3 问题求解

贝叶斯网络推理是概率分布的计算过程，即在给定模型中计算目标变量的后验概率。根据本文的多模块贝叶斯网络，可以从观测的事件出发逐层推理，得到置换节点的状态，最后推理得到恐怖袭击威胁等级。本文的多模块贝叶斯网络差异较大，因此针对不同的模块，本文选择不同的推理算法。对于分级贝叶斯网络模块，本文采用多树传播推理算法；其他 3 个贝叶斯网络模块采用联结树推理算法。联结树算法将有向图转换为树，从而大大减小了计算的复杂度，计算速度较快并且适用性强，对网络中存在多个查询节点进行推理时非常便捷。本文先将贝叶斯网络转化为联结树，然后对联结树上的消息传递过程进行定义，再计算概率，基本过程如下：

1、建立 Moral 图：将有向图转换为无向图，找出每个节点的父节点并连接起来把所有的有向边改为无向边

2、将 Moral 图三角化：三角化是指让图中不存在超过 3 个点的环。将大于等于 4 的环，取其中两个非相邻节点用无向边连接起来，进行 Moral 图的三角化，

即对顶点的一一删除。具体规则为（1）成对的联结该顶点的相邻节点；（2）删除该节点，并添加边；（3）删除的顶点必须保证，所需要添加的边最少。□□

3、在三角化的图中确定团：每个三角都代表了一个节点，两个相临的三角具有共同的边。这条边就成为两个节点之间的中间节点。如此就组成一张联通图。找到三角化后 Moral 图中构成联结树的所有团，团即为最大全联通子图，其中每对不同节点都有边相连。

5. 问题二

6. 问题三

7. 问题四

参考文献

附录 A 我的 MATLAB 源程序

```
while ~isempty(V)
[tmpd,j]=min(W(i,V));tmpj=V(j);
for k=2:ndd
[tmp1,jj]=min(dd(1,k)+W(dd(2,k),V));
tmp2=V(jj);tt(k-1,:)=[tmp1,tmp2,jj];
end
tmp=[tmpd,tmpj,j;tt];[tmp3,tmp4]=min(tmp(:,1));
if tmp3==tmpd, ss(1:2,kk)=[i;tmp(tmp4,2)];
else,tmp5=find(ss(:,tmp4)~=0);tmp6=length(tmp5);
if dd(2,tmp4)==ss(tmp6,tmp4)
ss(1:tmp6+1,kk)=[ss(tmp5,tmp4);tmp(tmp4,2)];
else, ss(1:3,kk)=[i;dd(2,tmp4);tmp(tmp4,2)];
end;end
dd=[dd,[tmp3;tmp(tmp4,2)]];V(tmp(tmp4,3))=[];
[mdd,ndd]=size(dd);kk=kk+1;
end; S=ss; D=dd(1,:);
```