“华为杯”第十五届中国研究生数学建模竞赛

学　　校	空军工程大学
参赛队号	**18900450036**
队员姓名	**1. 付其喜**
	2. 范翔宇
	3. 胡利平

“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

本文旨在基于全球恐怖主义数据库（GTD）所收集的最全面的恐怖袭击活动数据基础之上，通过大数据分析进行恐怖主义形成原因整理、时空蔓延特性监理与级别指数分布梳理，深入挖掘并量化评估各恐怖主义事件信息，克服恐怖事件不确定性、恐怖主义无边界性的难题，推测各恐怖主义组织与对应事件的相关度，并以此为牵引研判未来反恐趋势。最后筛选出与我国社会情况相似的区域，研究彼此共通性与差异性，结合量化与质化数据分析，探讨我国反恐维稳可鉴之策。

针对问题一，根据待评估事件数据的完备性，分别建立基于 RS-TOPSIS 与 ITR-RS（Rough Set）的危害程度量化评估模型。针对完备信息条件，本文提出 RS-TOPSIS 模型可以基于数据驱动，实现对事件的定量评估。针对不完备信息情况，提出了 ITR-RS 修正属性权重实现更为准确的评估。利用上述两种方法根据数据完备情况分别对表 1 事件与附件 1 中的所有事件进行量化评估与定级。表 1 事件定级为 200108110012（5 级）、200511180002（5 级）、200901170021（1 级）、201402110015（3 级）、201405010071（3 级）、201411070002（2 级）、201412160041（5 级）、201508010015（1 级）、201705080012（5 级）。同时，考虑到同样的参数在不同的年份中的作用影响不一样，因而不能将所有年份的数据放在一起进行评估，为此需要利用上述两种模型逐年评估，得到每年排在前十位的恐怖袭击事件，再构建相对危害值，削弱年份的影响，将 20 年中两种方法分别评出的每年前十事件，共计 400 组事件进行排序，获取前十位，即为十大恐怖袭击事件，分别为 200109110004、200409010002、201406150063、201710140002、200403210001、201607020002、199808070002、201404150089、200811260006、200403110004。

针对问题二，虽然可供学习的样本总量较多，可部分已知袭击者的事件样本不足十个甚至更低，为此，本文对有限的有标签样本进行半监督学习，提升小样本学习的准确性。为提升学习精度或速度，本文分别提出了 IS3VM（Improved Semi-supervised Support Vector Machines）与 RSNB（Revised Semi-supervised Naïve Bayes）算法，通过对比，效果优于改进前和其他的主流方法。从而训练好网络，将不确定实施者的事件输入，

即可锁定恐怖组织。将 2015、2016 两年制造的恐怖袭击事件危害值累加，得到排在前五的组织。为了确定前五位组织与某些事件的联系，利用五个组织涉及的事件训练 SVM，再将待检测事件重复输入网络 100 次，根据输出五个组织的频次来量化嫌疑度。该方法虽然准确但存在冗余计算、长时损耗的缺点，为此，本文又利用类平均聚类的方法，将十组事件与五个组织的所有事件进行聚类分析，根据聚类结果的亲疏程度，确定各个组织的嫌疑度。最后再用构建的 IS3VM 对 2017 年的数据进行学习，得到了十组事件最可能的嫌疑人，明确三组事件很可能是 ISIL 所为，三组事件很可能是 Taliban 所为，其余四组可能不是五个组织所为，三组很可能为 SPLM-IO，一组很可能为 Seleka。

针对问题三，在研究主要原因时，利用简单高效的粗糙集理论，去除冗余属性，根据指标重要程度，确定主要影响因素。为研究蔓延特性，借鉴传染病 SIR（Susceptible Infected Recovered）模型，对恐怖袭击事件的发展趋势进行推断。选定某个恐怖主义组织，认为其时空特性与级别分布可以考虑成时序序列，进而利用 RBF（Radical Basis Function）网络训练时序预测模型。而由于 RBF 网络需要确定输出时序的数量，为此，可通过蔓延特性确定某段时间内发生恐怖袭击事件的总数，进而完成时序模型的预测。将某块区域多个恐怖主义组织的预测结果进行整合，对此区域的恐怖主义态势具有一定的代表性，本文通过此种方法预测了 2018 年全球与部分区域的恐怖主义态势，模型结果表明：国际社会所面临的恐怖主义威胁仍呈现曲折上升态势，国际反恐之路任重而道远。

针对问题四，考虑到有一类恐怖袭击事件，其并不发生在战火纷飞的区域，同时发生该事件的区域发生恐怖袭击事件的频率也不高。而此类区域一般为政权稳定，社会和谐的区域或国家，这与中国的社会现状极其相似。此类数据虽然不多却对我国具有指导意义，采用离群点检测的方法，确定出这些离群点。由于这些点所处的区域与社会状态，与我国现阶段更为贴近，因此更需要研究此类和平暴恐袭击事件，并将其进行数据挖掘与泛化分析，研究此类点对应的恐怖袭击事件，分析其数据，最终针对性的给出对于加强我国的反恐工作的意见与建议。

关键词： 恐怖袭击事件；信息完备性；量化评估；危害相对值；半监督学习 SVM；类平均聚类法；RBF 时间序列模型；离群点检测

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，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。本题主要讨论的是，立足全球恐怖主义数据库中1998-2017 年世界上发生的恐怖袭击事件的记录，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，并依据历史数据，对未知恐怖袭击事件负责人进行嫌疑确定，以及预判未来反恐态势，为应对恐怖袭击提供决策和方案支持。

1.2 问题描述

需要通过建立数学模型，解决以下几个问题：

问题一：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。

问题二：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上述可能性，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，并将结果填入表 2。

问题三：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。

问题四：你们认为通过数学建模还可以发挥附件 1 数据的哪些作用？给出你们的模型和方法。

2 合理假设及符号系统

2.1 问题假设

（1）假设同一个犯罪组织/个人的犯罪模式不会产生突变，即，该犯罪组织/个人的犯罪手法是一个确定的有限集合；

（2）假设每一年不会产生性质极其恶劣，危害程度极高的恐怖袭击事件不会超过 10 个，即，近二十年来危害程度最高的十大恐怖袭击事件不会在某一年集中爆发；

（3）假设，某一类事件出现频率较多时，该事件的某个属性对应的参数分布概率可用同类事件中该参数出现的频率近似表示；

（4）假设恐怖袭击事件已知类别，其所有属性相互独立；

（5）假设某些重点区域的恐怖袭击事件，绝大多数是由盘踞在该区域影响力较大的恐怖组织发起的。

2.2 符号系统

本文用到的符号及其意义如表 1 所示。

表 1 本文用到的部分符号说明

符号	说明
T	目标属性决策矩阵
S	信息系统
U	对象集
A	属性集
R_B	U 上的等价关系

注：为方便论文阅读，模型涉及到的部分参量符号未在表中一一列举，而是在各表达式后进行解释说明。

3 问题一的建模与求解

3.1 问题分析

为实现对于恐怖袭击事件的量化评估。本文将其分解为两个子问题进行研究。

第一个子问题是根据信息是否完备，构建不同的量化评估模型。浏览附件 1 中的数据，可以看出，只有部分事件的数据是完备的，而另一部分事件，例如 200901170021，201712010012 等约 4000 多次诸多事件，可能由于其处于动荡的区域，恐怖事件频发，或统计手段落后，数据处理能力较弱，导致危害性中的重要评估指标人员伤亡人数（nkill）没有记录，这样对于事件的评估影响较大。同时财产损失（property）中记录为-9，即“未知”不知道事件是否造成财产损失，大约有 21200 个事件。这个重要的评估指标的数据缺失，也会对评估产生重要影响。由于各个区域参数统计体系并不健全与统一，导致事件部分参数数据缺失的情况时有发生。同时，虽然记录的信息都有用，可部分信息是彼此重复的。因此，可以采用粗糙集（Rough Set, RS）的方式确定评估中的核心属性与冗余属性，进行危害程度的判定。并将量化评估分为完备信息条件下的量化评估，与不完备信息系统条件下的量化评估，最后将两种评估结果进行结合。即任务 1 要研究的子问题一的（A），（B）两个问题。

另一个子问题是时间跨度对于事件评估的影响。任务 1 要求列出近二十年来危害程度最高的十大恐怖袭击事件，而二十年时间跨度很长，必须要考虑每个数据的时间特性。例如某个恐怖事件造成 100 万的经济损失，在 1998 年可能是极其恶性的事件，而在 2008 年是可以在短时间内恢复的创伤，在 2017 年这个量级甚至可能不值一提。即使是完全相同的数据，放在不同的年份中，其危害程度是不一样的。因此，可以将每一年的恐怖事件放在一起进行评估，而不能将二十年所有的恐怖事件放在一起进行评估。为列举出近二十年来危害程度最高的十大恐怖袭击事件。采用子问题一中构建的量化评估模型对 20 年中每一年的所有事件进行评估，得到每一年中危害程度最高的 20 起事件，共计得到 400 组事件，再结合时间特性，研究这 400 组事件的危害程度排序，取排序结果中的前 10 位，为近二十年来危害程度最高的十大恐怖袭击事件。而如何去描述时间特性，且这个不同时间对危害评估产生什么样的影响，是本任务子问题二重点要研究的内容。

3.2 模型建立

为实现对 20 年的各个数据进行量化评估与分级，并罗列出近二十年来危害程度最高的十大恐怖袭击事件。本团队对此问题的求解流程如 3.1 所示。

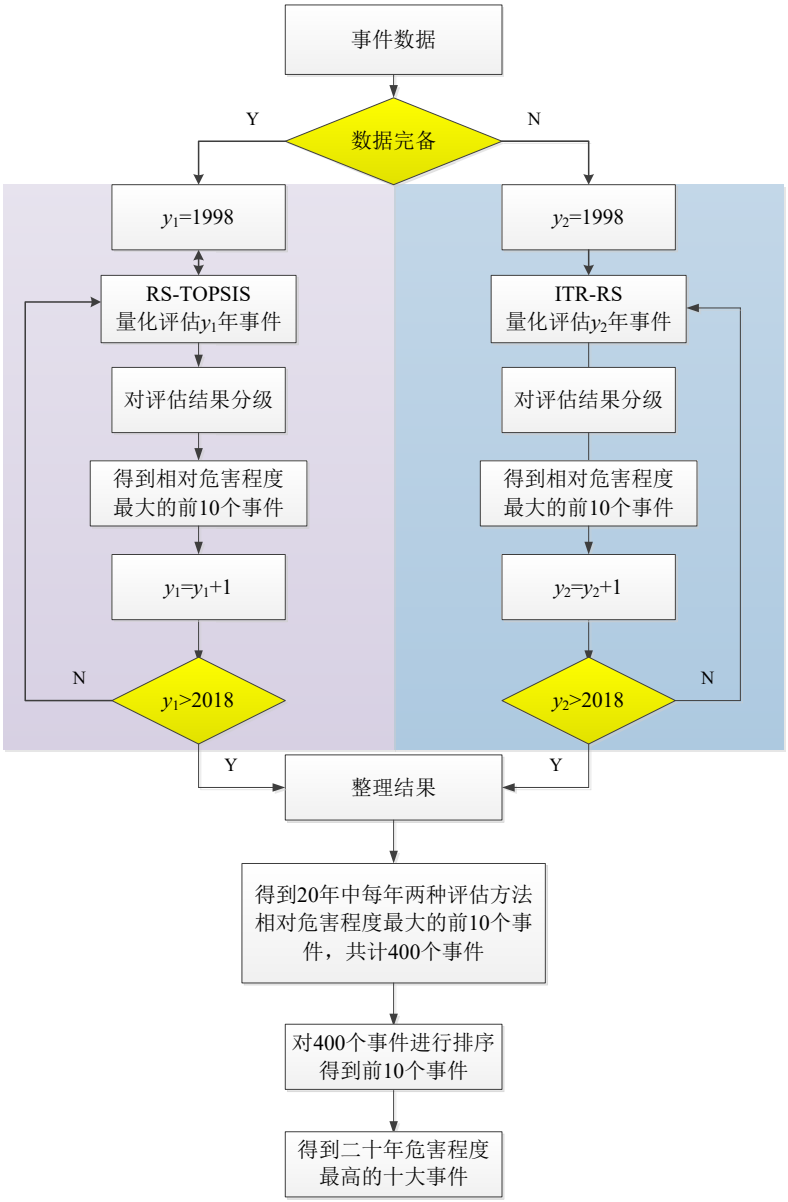


图 3.1 任务 1 求解思路

上述流程可以归结如下的 11 个步骤：

Step 1：首先得到原始数据，并对数据根据成本性、效益型、固定型以及区间型定量指标进行规范化，详见 3.2.1.2 节中 Step 2。预处理的方式在逐个判断所有的待评估的事件信息是否完备。判别方法会在 3.2.1.1 与 3.2.2.1 中论述。当事件的信息完备时，对于完备信息处理时转到 Step 2，不完备信息处理转至 Step 6。

Step 2：考虑到年份对评估的影响，因此，将评估逐年进行，初始化年份 y_1 为 1998 年。

Step 3：利用 RS-TOPSIS 方法对 y_1 年的事件进行量化评估，该方法会在 3.2.1 中介绍。并对评估结果进行进行分级。

Step 4：根据 RS-TOPSIS 方法评估结果，得到此种方法下 y_1 年的前 10 个相对本年危害程度最大的事件 $E_{\text{RS-TOPSIS}}(y_1,1),\ldots, E_{\text{RS-TOPSIS}}(y_1,10)$，并得到相对危害值，该方法会在 3.2.3 中介绍。

Step 5：年份 y_1 加 1，即完成本年的事件评估。判断新 y_1 是否大于 2018，即是否完

成 20 年的评判。不满足，返回 Step 3，满足则跳转至 Step 10。

Step6：初始化年份 y_2 为 1998 年。

Step7：利用 ITR-RS 方法对 y_2 年的事件进行量化评估，该方法会在 3.2.2 中介绍。并对评估结果进行进行分级。

Step8：根据 ITR-RS 方法评估结果，得到此种方法下 y_2 年的前 10 个相对危害程度最大的事件 $E_{\text{ITR-RS}}(y_2,1),\ldots, E_{\text{ITR-RS}}(y_2,10)$ ，并得到相对危害值，该方法会在 3.2.3 中介绍。

Step9：年份 y_2 加 1，即完成本年的事件评估。判断新 y_2 是否大于 2018，即是否完成 20 年的评判。不满足，返回 Step 7，满足则跳转至 Step 10。

Step10：根据年份，将两种方法的评估结果进行整理，得到事件的威胁度量值与等级，并得到得到 20 年中每年两种评估方法相对危害程度最大的前 10 个事件，共计 400 个事件。

Step11：根据上述 400 个事件，根据事件在当年的相对危害值进行排序，得到前 10 个事件，即为得到二十年危害程度最高的十大事件。

通过上述流程即可实现对于所有事件的危害程度量化与分类，并取得近 20 年危害程度最大的十大事件。

3.2.1 子问题 1（A）模型

根据图 3.1 中的论述，首先介绍如何确定对应事件数据是否完备，为此，本组借用粗糙集中完备信息系统与不完备信息系统的概念，对数据是否完备进行说明。

粗糙集理论算法简单，调用的计算资源较少，保证系统有闲置资源用于冗余设计，进而保证其鲁棒性。粗糙集处理易于实现，不需耗费大量的系统资源，只是进行简单的运算即可，算法的实时性可以得到保障。粗糙集理论作为一种数学理论，具有良好的数学基础作为支撑，大幅度提升了算法结果的可信度，保证了算法的准确性。为此，本组基于粗糙集理论，实现对于恐怖事件的量化评估与危害程度排序。

3.2.1.1 粗糙集基础

1． 粗糙集的基本概念

定义 1 称$\{\boldsymbol{U},\boldsymbol{A},\boldsymbol{F},\boldsymbol{d}\}$是决策信息系统，其中 $\boldsymbol{U}=\{x_1,x_2,\ldots, x_n\}$为对象集，$\boldsymbol{U}$ 中的每个元素 x_n（$i\le n$）称为一个对象。$\boldsymbol{A}=\{a_1,a_2,\ldots, a_m\}$为属性集，$\boldsymbol{A}$ 中的每个元素 a_l（$l\le m$）称为一个属性。$\boldsymbol{F}=\{f_l:\boldsymbol{U}\rightarrow\boldsymbol{V}_l$ （$l\le m$）$\}$为 $\boldsymbol{U}$ 与 $\boldsymbol{A}$ 之间的关系集，其中 $\boldsymbol{V}_l$ 为 a_l（$l\le m$）的值域。$\boldsymbol{d}:\boldsymbol{U}\rightarrow\boldsymbol{V}_d$ 为决策，$\boldsymbol{V}_d$ 取有限值。每个属性子集 $a\subseteq\boldsymbol{A}$ 决定了一个不可区分的关系 ind（$\boldsymbol{A}$）：

$$ind(\boldsymbol{A})=\{(x,y)\in\boldsymbol{U}*\boldsymbol{U}\mid\forall a\in\boldsymbol{A},a(x)=a(y)\} \tag{3.1}$$

关系 $ind(a)(a\subseteq\boldsymbol{A})$构成了 $\boldsymbol{U}$ 的划分，用 $\boldsymbol{U}/ind(a)$表示。

定义 2 称设$\{\boldsymbol{U},\boldsymbol{A},\boldsymbol{F},\boldsymbol{d}\}$是一个信息系统，对于任意 $\boldsymbol{B}\subseteq\boldsymbol{A}$，记：

$$\boldsymbol{R}_{\boldsymbol{B}}=\{(x_i,x_j)\mid f_l(x_i)=f_l(x_i)(a_l\in\boldsymbol{B})\} \tag{3.2}$$

$\boldsymbol{R}_{\boldsymbol{B}}$ 则是 $\boldsymbol{U}$ 上的等价关系，记：

$$[x_i]_{\boldsymbol{B}}=\{x_j\mid(x_i,x_i)\in\boldsymbol{R}_{\boldsymbol{B}}\} \tag{3.3}$$

则 $\boldsymbol{U}/\boldsymbol{R}_{\boldsymbol{B}}=\{[x_i]_{\boldsymbol{B}}|x_i\in\boldsymbol{U}\}$是 $\boldsymbol{U}$ 上的划分。同理：

$$\boldsymbol{R}_{\boldsymbol{d}}=\{(x_i,x_j)\,|\,d(x_i)=d(x_j)\} \tag{3.4}$$

2. 属性约简流程

Step1：构造决策辨识集

设$\{\boldsymbol{U},\boldsymbol{A},\boldsymbol{F},\boldsymbol{d}\}$为决策信息系统，记：

$$\boldsymbol{U}/\boldsymbol{R}_{\boldsymbol{A}}=\{[x_i]_{\boldsymbol{A}}\,|\,x_i\in\boldsymbol{U}\} \tag{3.5}$$

$$\boldsymbol{U}/\boldsymbol{R}_{\boldsymbol{d}}=\{[x_i]_{\boldsymbol{d}}\,|\,x_i\in\boldsymbol{U}\} \tag{3.6}$$

$$\boldsymbol{D}_{\boldsymbol{d}}([x_i]_{\boldsymbol{A}},[x_j]_{\boldsymbol{A}})=\begin{cases}\{a_l\in\boldsymbol{A}\,|\,f_l(x_i)\neq f_l(x_i)\}, & [x_i]_{\boldsymbol{d}}\cap[x_j]_{\boldsymbol{d}}=\varnothing\\ \varnothing, & [x_i]_{\boldsymbol{d}}\cap[x_j]_{\boldsymbol{d}}\neq\varnothing\end{cases} \tag{3.7}$$

称 $\boldsymbol{D}_{\boldsymbol{d}}([x_i]_{\boldsymbol{A}},\ [x_j]_{\boldsymbol{A}})$为$[x_i]_{\boldsymbol{A}}$ 与$[x_j]_{\boldsymbol{A}}$ 的决策辨识集，称：

$$\boldsymbol{D}_{\boldsymbol{d}}=(\boldsymbol{D}_{\boldsymbol{d}}([x_i]_{\boldsymbol{A}},[x_j]_{\boldsymbol{A}})\,|\,[x_i]_{\boldsymbol{A}},[x_j]_{\boldsymbol{A}}\in\boldsymbol{U}/\boldsymbol{R}_{\boldsymbol{A}}) \tag{3.8}$$

为决策信息系统的决策辨识矩阵。

Step2：构造决策约简集

对于决策信息系统，若 $\boldsymbol{B}$ 为决策协调集，当且仅当对于任意的 $\boldsymbol{D}_{\boldsymbol{d}}([x_i]_{\boldsymbol{A}},\ [x_j]_{\boldsymbol{A}})\neq\varnothing$，有

$$\boldsymbol{B}\cap\boldsymbol{D}_{\boldsymbol{d}}([x_i]_{\boldsymbol{A}},[x_j]_{\boldsymbol{A}})\neq\varnothing \tag{3.9}$$

且 $\boldsymbol{B}$ 的任何真子集均不为决策协调集时，称 $\boldsymbol{B}$ 为决策约简集。即可以保留系统决策不变的约简属性集，属性约简后可以降低系统的冗余度。

3. 权重确定流程

Step1：构建条件属性 A 中的各个属性 ai 与决策属性 d 关于论域的分类，得到：$\boldsymbol{U}/ind(a_i)$，i=1,2,…,n 与 $\boldsymbol{U}/ind(\boldsymbol{d})$。

Step2：依次去掉各条件属性，得到新的分类：

$\boldsymbol{U}/ind(\boldsymbol{A}\text{-}a_i)$，$i$=1,2,…,$n$。

Step3：利用下式，计算决策属性 $\boldsymbol{d}$ 对条件属性 $\boldsymbol{A}$ 的支持度（依赖程度）：

$$K=I_{\boldsymbol{A}}(\boldsymbol{d})=\frac{|\,POS_{\boldsymbol{A}}(\boldsymbol{d})\,|}{|\,\boldsymbol{U}\,|} \tag{3.10}$$

其中$|\boldsymbol{U}|$和$|POS_{\boldsymbol{A}}(\boldsymbol{d})|$分别为论域和正域的基数，即其中包含元素的个数。正域$|POS_{\boldsymbol{A}}(\boldsymbol{d})|$表示那些根据属性知识判定肯定属于 x 中的元素所组成的最大集合：

$$\underline{\boldsymbol{R}}X=\{x\in\boldsymbol{U}\,|\,[x]_{\boldsymbol{R}}\subseteq\boldsymbol{U}\} \tag{3.11}$$

Step4：利用（3.10）式，计算决策属性 $\boldsymbol{d}$ 对去掉某个属性后的分类$(\boldsymbol{A}\text{-}a_i)$的支持度；

Step5：利用（3.11）式，计算条件属性关于决策属性的重要性：

$$SGF_{A-a_i}=I_{\boldsymbol{A}}(\boldsymbol{d})-I_{\boldsymbol{A}-a_i}(\boldsymbol{d}) \tag{3.12}$$

Step6：可得条件属性的权重为：

$$\omega(a_i)=\frac{SGF_{A-a_i}}{\sum_{i=1}^{n}SGF_{A-a_i}} \tag{3.13}$$

通过上述流程即可完成对于属性权重的确定。

3.2.1.2 TOPSIS 恐怖事件的量化评估与危害程度排序算法

恐怖事件的量化评估与危害程度排序是基于每个事件的参数进行的，即利用多维参数判定输出目标。因而可将转换为多属性决策问题。可以利用 TOPSIS 法进行恐怖事件的量化评估与危害程度排序。其具体步骤如下：

Step1：根据数据库，得到某个恐怖事件的全部数据，构建目标属性决策矩阵 $\boldsymbol{T}=(v_{ij})_{m\times n}$，$v_{ij}$ 表示第 i 个事件（样本）第 j 个数据属性的指标值，如下式所示：

$$\boldsymbol{T}=\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{bmatrix}_{m\times n} \tag{3.14}$$

Step2：对成本性、效益型、固定型以及区间型定量指标进行规范化。效益型就是指其值越大，目标函数函数越大，计算公式如下：

$$\gamma_{ij}=\frac{v_{ij}-\wedge v_{ij}}{\vee v_{ij}-\wedge v_{ij}} \tag{3.15}$$

成本型就是指其值越大，目标函数函数越小，计算公式如下：

$$\gamma_{ij}=\frac{\vee v_{ij}-v_{ij}}{\vee v_{ij}-\wedge v_{ij}} \tag{3.16}$$

其中，“∧”表示合取运算，“∨”表示析取运算。

固定型是指其值既不能太大，又不能太小，而以稳定在某个固定值为最佳的指标；或者说，其值越接近某个值越好的指标。

区间型是指其值以落在某个固定区间为最佳的指标，或者说，其值越接近某个固定区间（包括落入该区间）越好的指标，象国标中规定的等级划分通常属于此类型指标。

基于公式（3.15）和（3.16）以及固定型与区间型对附件 1 中的数据进行简要处理。得到规范化评价矩阵 $\boldsymbol{V}=(\tilde{v}_{ij})_{m\times n}$：

$$\boldsymbol{V}=\begin{bmatrix} \tilde{v}_{11} & \tilde{v}_{12} & \cdots & \tilde{v}_{1n} \\ \tilde{v}_{21} & \tilde{v}_{22} & \cdots & \tilde{v}_{2n} \\ \vdots & \vdots & & \vdots \\ \tilde{v}_{m1} & \tilde{v}_{m2} & \cdots & \tilde{v}_{mn} \end{bmatrix}_{m\times n} \tag{3.17}$$

Step3：对规范化评价矩阵 $\boldsymbol{V}$ 进行加权处理，得到加权标准化的决策矩阵：

$$\boldsymbol{C}=(c_{ij})_{m\times n}=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix}_{m\times n}=(\omega_j\tilde{v}_{ij})_{m\times n} \tag{3.18}$$

Step4：确定理想解 $\boldsymbol{C}^{+}$与负理想解 $\boldsymbol{C}^{-}$，理想解为每个评价指标下各个样本中危害值最大的解，负理想解则为危害值最小的解：

$$\boldsymbol{C}^{+}=\{\max_{1\le i\le m} c_{ij}\}=\{c_1^{+},c_2^{+},\cdots,c_n^{+}\} \tag{3.19}$$

$$\boldsymbol{C}^{-}=\{\min_{1\le i\le m} c_{ij}\}=\{c_1^{-},c_2^{-},\cdots,c_n^{-}\} \tag{3.20}$$

Step5：计算各个样本到正、负理想解的距离 S_i^{+} 、 S_i^{-}：

$$S_i^{+}=\sqrt{(\sum_{j=1}^{n} c_{ij}-c_j^{+})^2} \tag{3.21}$$

$$S_i^{-}=\sqrt{(\sum_{j=1}^{n} c_{ij}-c_j^{-})^2} \tag{3.22}$$

Step6：计算各个目标相对于 $\boldsymbol{C}^{+}$的相对贴近度 CL_i：

$$CL_i=\frac{S_i^{-}}{S_i^{-}+S_i^{+}} \tag{3.23}$$

Step7：相对贴近度即该每个对象的危害度量值，形成事件的危害队列矢量 $\boldsymbol{T_e}$，可以表示为：

$$\boldsymbol{T_e}=[CL_1,...CL_i,...CL_m] \tag{3.24}$$

根据危害数值大小，对危害队列 $\boldsymbol{T_e}$ 进行降序排列，则得到危害评估结果 $\boldsymbol{T_e}'$：

$$\begin{aligned} \boldsymbol{T_e}' &= sort[\boldsymbol{T_e}]=sort([CL_1,...,CL_i,...,CL_m]) \\ &=[T'_{e\,1},...,T'_{e\,j},...,T'_{e\,m}] \end{aligned} \tag{3.25}$$

$$T_{ej}'=CL_i,\ i,j\in[1,m],s.t\ \ T_{ej}'>T_{ej+1}'\ \ j\in[1,m] \tag{3.26}$$

3.2.1.3 基于 RS-TOPSIS 算法的恐怖事件的量化评估与危害程度排序流程

在上述处理流程 Step3 中的权重系数一般由人为给定，由专家给定的权重具有一定作用，但经验源于过往的实践，难以保证之前的积累在现今的条件下能否适用，同时不同年代的恐怖袭击时间的权重必然会发生变化。为此，本文采用粗糙集理论，基于每一年的数据，规避主观因素的影响与对先验信息的需求，确定 TOPSIS 方法中的权重，拓宽其算法的适用性，从而构建基于 RS-TOPSIS 方法的恐怖事件的量化评估与危害程度排序模型，用以度量恐怖事件的危害程度。综上所述，基于 RS-TOPSIS 的恐怖事件的量化评估与危害程度排序流程如图 3.2 所示。

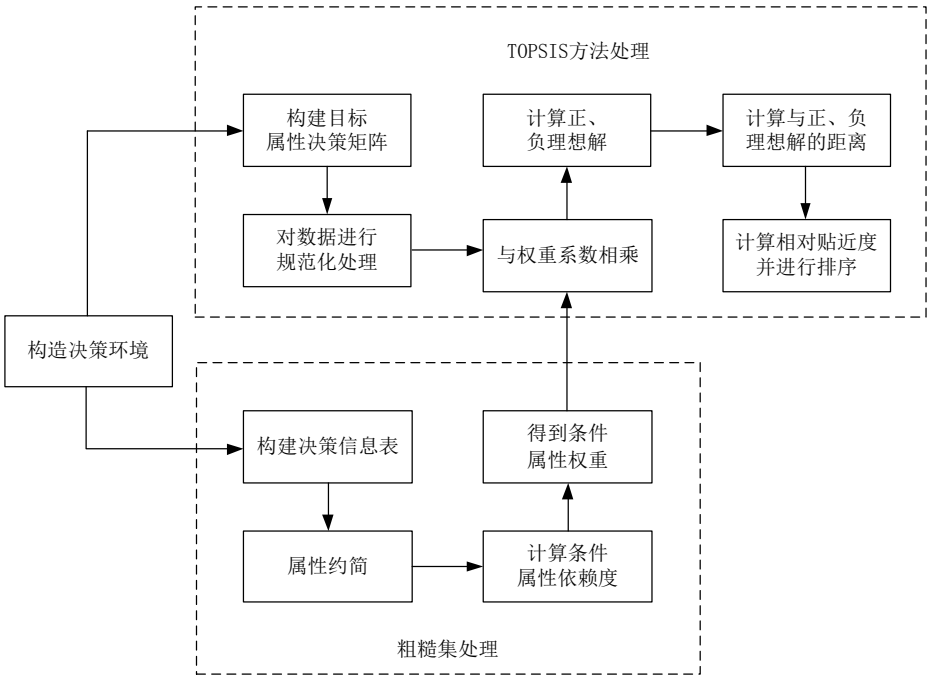


图 3.2　危害评估处理流程

根据数据库得到的每个事件对应的机载电子对抗设备将测得 PDW 输入威胁评估流程,将初始形成的数据集用来构建 TOPSIS 的目标属性矩阵与粗糙集处理所需的信息表，即对数据进行双路并行处理；一路基于粗糙集理论，将数据离散化后进行属性约简，并计算条件属性的依赖度，从而得到条件属性的权重；另一路采用 TOPSIS 处理方式，将决策矩阵进行规范化处理后，与粗糙集处理得到的权重系数相乘，基于得到的新数据计算正、负理想解和样本与它们之间的距离，最后计算相对贴近度，即为事件危害程度的量化指标。

上述流程判定的危害程度度量值介于[0,1]之间，将其等分为 5 段，即[0,0.2]为一级，[0.2,0.4]为二级，[0.4,0.6]为三级，[0.6,0.8]为四级，[0.8,1]为五级，进而完成对恐怖事件进行分级。

本节的研究表明，二者的结合使用能有效对恐怖事件的危害程度进行判定，从而为后续的评估甚至反恐提供理论与参数依据。TOPSIS 方法的过程能够固化，且计算简便，粗糙集基于数据驱动，降低了主观因素的影响，将两种方法进行有机融合，对于先验信息的需求程度降低，为恐怖事件的量化评估与危害程度排序提供了一种新的工程决策方法。

3.2.2 子问题 1（B）模型

很明显，附件 1 中存在部分数据的缺失，为了能够进一步度量事件的威化成都，需要对其数据进行处理。因此，本节重点研究在不完备信息系统下，事件的危害程度量化评估方法。

3.2.2.1 不完备信息系统

设 $S=\{U,AT,V,F\}$ 为信息系统，其中，$\boldsymbol{AT}=\boldsymbol{A}\cup\boldsymbol{d}$，条件属性 $\boldsymbol{A}=\{a_1, a_2, \ldots, a_m\}$ 的集合中存在属性值缺失，此时，$\boldsymbol{S}$ 为不完备决策信息系统。若每个属性子集 $a\subseteq\boldsymbol{A}$，可

以确定一个不可区分的关系 *ind*(***A***)，即：

$$ind(A)=\{(x,y)\in U*U\mid \forall a\in A,a(x)=a(y)\} \tag{3.27}$$

关系 *ind*(***A***)构成了 ***U*** 的划分，用 ***U***/*ind*(***A***)表示。

为进行模型构建，对不完备决策信息系统 ***S***={***U***, ***AT*** ,***V***, ***F***}需要明确下列基本概念：

定义 3：等价关系

对于任意 ***B***⊆***A***，记：

$$R_B=\{(x_i,x_j)\mid f_l(x_i)=f_l(x_i)(a_l\in B)\} \tag{3.28}$$

R_B 则 ***U*** 是上的等价关系，记：

$$[x_i]_B=\{x_j\mid (x_i,x_i)\in R_B\} \tag{3.29}$$

则 ***U***/R_B=$\{[x_i]_B|x_i\in \boldsymbol{U}\}$是 ***U*** 上的划分。

定义 4：容差关系

$$T(x,y)=\{(x,y)\in U^2\mid \forall a\in A(a(x)=*\vee a(y)=*\vee a(x)=a(y))\} \tag{3.30}$$

定义 5：决策辨识集

$$U/R_A=\{[x_i]_A\mid x_i\in U\} \tag{3.31}$$

$$U/R_d=\{[x_i]_d\mid x_i\in U\} \tag{3.32}$$

$$D_d([x_i]_{AT},[x_j]_{AT})=\begin{cases}\{a_l\in A_T\mid f_l(x_i)\neq f_l(x_j)\},[x_i]_d\bigcap[x_j]_d\neq\varnothing\\ \varnothing\qquad\quad ,[x_i]_d\bigcap[x_j]_d=\varnothing\end{cases} \tag{3.33}$$

称 $D_d([x_i]_A,[x_j]_A)$为$[x_i]_A$ 与$[x_j]_A$ 的决策辨识矩阵，记为：

$$D_d=(D_d([x_i]_A,[x_j]_A)\mid [x_i]_A,[x_j]_A\in U/R_A) \tag{3.34}$$

定义 6：决策约简集

设 ***B*** 为决策协调集，当且仅当对于任意的 $D_d([x_i]_A,[x_j]_A)\neq\varnothing$，有：

$$B\bigcap D_d([x_i]_A,[x_j]_A)\neq\varnothing \tag{3.35}$$

当 ***B*** 的任何真子集均不是决策协调集时，此时，称 ***B*** 为决策约简集，即能够保留系统决策不变的约简属性集，通过属性约简能够降低系统的冗余度[85]。

定义 7：广义决策函数

广义决策函数∂_A（x）可表示如下：

$$\partial_A(x)=\{i\mid i=d(y),y\in T_A(x)\} \tag{3.36}$$

定义 8：区分函数与约简计算

设 $\alpha_A(x,\ y)$是满足$(x,\ y)\notin$T($\{a\}$)的属性 $a\in$***A*** 的集合，因此，如果$(x,\ y)\in$T(***A***)，则 $\alpha_A(x,\ y)\neq\varnothing$。令$\sum\alpha_A(x,\ y)$是一个布尔表达式，如果 $\alpha_A(x,\ y)\neq\varnothing$，则$\sum\alpha_A(x,\ y)$=1；否则，$\sum\alpha_A(x,\ y)$是包含在 $\alpha_A(x,y)$中的属性所对应变量的析取。

Δ^*是不完备决策表中的区分函数，若：

$$\Delta^* = \prod_{(x,y)\in U\{z\in U|d(z)\notin\partial_{AT}(x)\}} \sum \alpha_{AT}(x,y) \tag{3.37}$$

$\Delta^*(x)$是不完备决策表中对象 x 的区分函数，若：

$$\Delta^*(x) = \prod_{y\in U\{z\in U|d(z)\notin\partial_{AT}(x)\}} \sum \alpha_{AT}(x,y) \tag{3.38}$$

3.2.2.2 ITR-RS 算法

（1）ITR-RS 算法

根据定义 3，在容差关系粗糙集中，未知值“*”可以与任何已知的属性值相等，其划分粒度过大，很容易造成两个没有明确相同已知属性值的个体被划分到同一个容差类中，不便在实际中应用。针对传统容差关系粗糙集的局限，本节提出一种改进的容差关系，在不引入信息系统外知识的情况下考虑了属性的权重，更符合实际应用情况。

在完备决策信息系统 $\boldsymbol{S}=\{\boldsymbol{U},\ \boldsymbol{AT}\ ,\boldsymbol{V},\ \boldsymbol{F}\}$中，$\boldsymbol{B}\subseteq\boldsymbol{A}$，$0\leqslant\omega(b)\leqslant 1$ 为属性 b 在 $\boldsymbol{B}$ 中的权重，定义加权阈值容差关系为：

$$\begin{aligned} WT(\omega) = \{(x,y)\,|\, x\in U\wedge y\in U\wedge\forall b\in B(b(x)= \\ *\vee b(y)=*\vee b(x)=b(y))\wedge\sum_{b\in B}\omega(b)\geq\omega\} \end{aligned} \tag{3.39}$$

其中，

$$B' = \{b\in B(b(x)\neq *)\wedge b(y)\neq *)\wedge(b(x)=b(y))\},0\leq\omega\leq 1 \tag{3.40}$$

B'为阈值，此时，记$[x]_B^{WT(\omega)} = \{y\in U\,|\,(x,\ y)\in WT(\omega)\}$为对象 x 的加权阈值容差类。

由（3.39）式可知，对象 x 与 y 只有在属性子集 $\boldsymbol{B}$ 中取值明确相同的属性权重和大于等于 ω，且不存在明确不相同的属性值时，才能被判定为同一类。在实际应用中，ω 的取值根据实际情况和人的主观要求得到。

在不完备信息系统 $\boldsymbol{S}=\{\boldsymbol{U},\ \boldsymbol{AT}\ ,\boldsymbol{V},\ \boldsymbol{F}\}$中，对象集 $\boldsymbol{X}$ 关于属性子集 $\boldsymbol{B}$ 的加权阈值容差关系的上近似 $B^{\mathrm{WT}(\omega)}(X)$和下近似 $\mathrm{B}_{\mathrm{WT}(\omega)}(X)$分别为：

$$B^{WT(\omega)}(X) = \{x\,|\,[x]_B^{WT(\omega)}\cap X\neq\varnothing, x\in U\} = \bigcup_{x\in X}[x]_B^{WT(\omega)} \tag{3.41}$$

$$B_{WT(\omega)}(X) = \{x\,|\,[x]_B^{WT(\omega)}\subseteq X, x\in U\} \tag{3.42}$$

（2）属性权重的确定

在改进的容差关系粗糙集模型中，根据不完备信息系统的信息量计算属性权重 ω(b)，具体求解步骤如下：

Step1：计算属性子集 $\boldsymbol{B}$ 的信息量。对不完备信息系统$\boldsymbol{S}=\{\boldsymbol{U},\ \boldsymbol{AT}\ ,\boldsymbol{V},\ \boldsymbol{F}\}$，$\boldsymbol{B}\subseteq\boldsymbol{A}$，$[x_i]_B^T$ 为 x_i 的容差类，根据（3.43）式计算信息量 $I(\boldsymbol{B})$：

$$I(B) = \frac{n}{n-1} - \frac{1}{n(n-1)}\sum_{i=1}^{n}|[x_i]_B^T| \tag{3.43}$$

这里，$0\leqslant I(\boldsymbol{B})\leqslant 1$，max $I(\boldsymbol{B})=1$，即，$[x_i]_B^T = \{x_i\}$；min $I(B)=0$，即，$[x_i]_B^T = U$。

Step2：计算属性重要程度。根据（3.44）式，计算属性 $b\in\boldsymbol{B}\subseteq\boldsymbol{A}$ 在 $\boldsymbol{B}$ 中的重要程度。

$$\mathrm{Sig}_B(b) = I(B) - I(B\setminus\{b\}) \tag{3.44}$$

这里，0≤Sig$_B$（b）≤1。

Step3：计算属性权重。根据（3.45）式，计算属性权重 ω(b)。

$$\omega(b) = \frac{Sig_B(b)}{\sum_{b_i \in B} Sig_B(b_i)} \tag{3.45}$$

这里，$0 \le \omega(b) \le 1, \sum_{b\in B} \omega(b) = 1$。

3.2.2.3 基于 ITR-RS 的恐怖事件的量化评估与危害程度排序流程

基于 ITR-RS 的不完备信息系统恐怖事件的量化评估与危害程度排序具体流程如下：

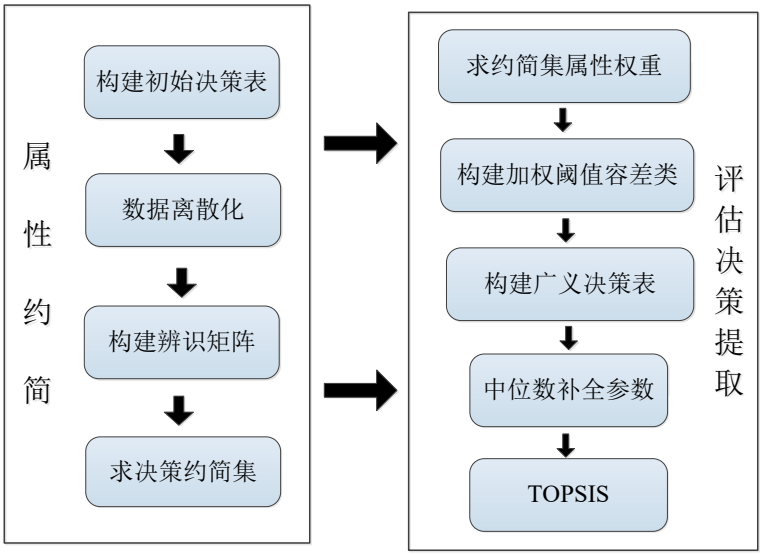


图 3.3 基于 ITR-RS 的不完备信息系统恐怖事件的量化评估与危害程度排序流程图

具体步骤如下：

Step1：构建初始决策表。确定不完备目标事件的信息系统 $\boldsymbol{S}$，确定目标集 $\boldsymbol{U}$={x_1, x_2, …, x_n}，属性指标集 $\boldsymbol{A}$= {a_1, a_2, …, a_m}和决策信息 $\boldsymbol{d}$；

Step2：数据离散化。根据等间隔法对属性信息离散化；

Step3：属性约简。根据定义 5 构建不完备决策信息系统的辨识矩阵，并根据（3.35）式得到系统的决策约简集，实现属性约简；

Step4：计算属性权重。根据（3.43）-（3.45）式计算约简后属性的权重；

Step5：构建广义决策表。根据（3.39）式计算 U 上的加权阈值容差类，并结合（3.41）、（3.42）式得到系统决策的上下近似，再根据（3.36）式构建广义决策表；

Step6：对缺失的数据进行中位数补全。

Step7：利用 TOPSIS 对不完备恐怖事件信息系统的量化评估与危害程度排序。

上述流程判定的危害程度度量值介于[0,1]之间，将其等分为 5 段，即[0,0.2]为一级，[0.2,0.4]为二级，[0.4,0.6]为三级，[0.6,0.8]为四级，[0.8,1]为五级，进而完成对恐怖事件

进行分级。

在将两种方法的结果结合，即可得到各个事件的评估分级。

3.2.3 子问题 2 模型

同样的数据放在不同的年份其作用意义是不一样。例如某个恐怖事件造成 100 万的经济损失，在 1998 年可能是极其恶性的事件，而在 2008 年是可以在短时间内恢复的创伤，在 2017 年这个量级甚至可能不值一提。即使是完全相同的数据，放在不同的年份中，其危害程度是不一样的。同样的数据放在不同的年份，评估结果应该会有所区别，

因此，为了得到近二十年来危害程度最高的十大恐怖袭击事件，不能将所有的恐怖事件不记年份的放在一起一并评估，而是要根据其年份逐年进行评估，再将结果转换，得到一个与时间关系不大的评估指标来进行统一度量。

为得到危害程度前十位的事件，本节采用危害相对度来衡量本年的恐怖事件的强度。以 RS-TOPSIS 评估方法为例，进行说明，流程图如图 3.4 所示。

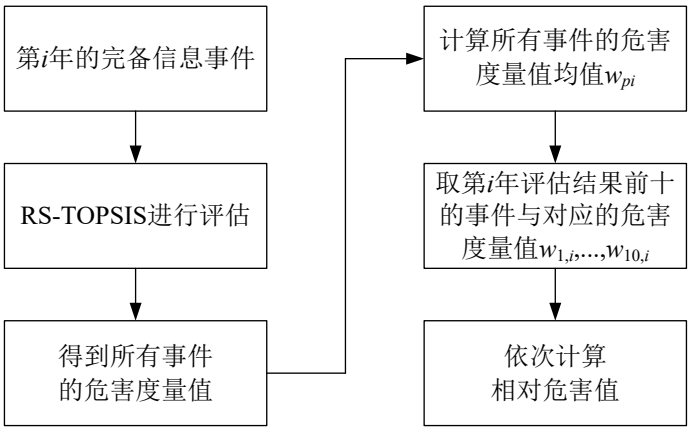


图 3.4 RS-TOPSIS 确定第 i 年十大危害程度最大事件流程

上述流程可以归结为六个步骤，详细为：

Step 1：取出第 i 年的数据完备的事件，假设共计 N_i 件。

Step 2：采用前文构建的 RS-TOPSIS 算法对第 i 年的 N_i 件完备事件进行危害度量，得到危害度量值；

Step 3：计算 Step 2 中 N_i 件恐怖事件的危害度量值均值 w_{pi}。

Step 4：取第 i 年评估结果前十的事件与对应的危害度量值 $w_{1,i},...,w_{10,i}$。

Step 5：利用下式，依次计算十个事件的相对危害值 $k_{j,i}, j \in [1,10]$。

$$k_{j,i}^{RS-TOPSIS} = \frac{w_{j,i} - w_{pi}}{w_{pi}} \tag{3.46}$$

通过上述流程，即可得到第 i 年当年的十大恐怖袭击事件的相对度量值。

利用 ITR-RS 确定第 i 年十大危害程度最大事件流程，只是在处理的数据与评估方法上有所差别，其他步骤雷同，最后计算方法为：

$$k_{j,i}^{ITR-RS} = \frac{w_{j,i} - w_{pi}}{w_{pi}} \tag{3.47}$$

通过 RS-TOPSIS 方法与 ITR-RS 方法，可以得到第 i 年的各自的十大危害事件，即：

$$k_i = \left[k_{1,i}^{RS-TOPSIS}, \ldots, k_{10,i}^{RS-TOPSIS}, k_{1,i}^{ITR-RS}, \ldots, k_{10,i}^{ITR-RS} \right] \tag{3.48}$$

得到上述 1*20 的矩阵，再将该方法对每一年的数据进行处理，得到如下的结果：

$$k = \begin{bmatrix} k_{1,1998}^{RS-TOPSIS}, \ldots, k_{10,1998}^{RS-TOPSIS} & , & k_{1,1998}^{ITR-RS}, \ldots, k_{10,1998}^{ITR-RS} \\ \cdots & \ddots & \cdots \\ k_{1,2017}^{RS-TOPSIS}, \ldots, k_{10,2017}^{RS-TOPSIS} & , & k_{1,2017}^{ITR-RS}, \ldots, k_{10,2017}^{ITR-RS} \end{bmatrix} \tag{3.49}$$

得到上述的 20*20 矩阵，即 20 年中，每一年的两种方法评估的一年内的危害成都最大的十个事件。这样共计 400 个事件，将这 400 个事件从大到小依次排序，根据 k 值取出排在前 10 位的事件，即为题目中所要求的近 20 年中危害程度最大的十大事件。

为得到影响最大的十大恐怖事件，根据本节前文所述，由于数据具有时代性，不能将所有的数据放在一起一并处理。因此，要得到一个与时间不太相关的量来度量某个事件对这 20 年的影响。上文的 k 值，从公式（3.47）可以看出，描述的是这个事件对当年的影响程度，如果某个事件是 20 年中影响巨大的恐怖袭击事件，其影响应该远高于当年的恐怖事件度量值均值。因此，其危害程度应远高于当年的危害程度均值，采用相对值的方法，能够大幅度降低时间所带来的影响，从而根据相对值这个指标，将 20 年的评估结果拉平，从而可以将这 400 组数据进行对比排序，进而得到近 20 年中危害程度最大的十大恐怖袭击事件。

3.3 模型解算及结果分析

3.3.1 子问题 1（A）求解

为保证处理方法的可信性与普适性，同时基于上文分析，评估是针对每一年进行的，本团队随机选取附件 1 中事件参数作为危害量化评估的仿真验证条件，由于篇幅所限，本节只是做示例性验证，为表达清晰，随机选取 2017 年的 10 个事件构建已知方案集，并选取 8 个形成条件属性，建立如表 3.1 所示的原始样本决策数据，得到样本集 $\boldsymbol{U}=\{x_1, x_2,\ldots,x_{10}\}$，指标（属性）集 $\boldsymbol{A}=\{a,b,c,d,e,f,g,h\}$，决策属性 D。

其中，x_1：201701010001；x_2：201702220014；

x_3：201703300002；x_4：201704070046；

x_5：201705250037；x_6：201706010006；

x_7：201707110063；x_8：201708280027；

x_9：201709060009；x_{10}：201710240028。

a：死亡总数（nkill）；

b：受伤总数（nwound）；

c：财产损失的价值（美元）（propvalue）；

d：武器类型（weapontype1）；

e：地区（region）；

f：攻击类型（attacktype1）；

g：持续天数（extended）；

h：财产损害程度（propextent）；

D：疑似恐怖主义（doubtterr）。

表 3.1　原始数据

对象	*D*	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
x_1	0	39	69	-99	5	10	2	0	4
x_2	1	15	19	-99	5	11	2	0	3
x_3	1	1	3	-99	5	6	2	0	3
x_4	1	0	2	-99	5	6	2	0	4
x_5	1	9	0	-99	6	10	3	0	4
x_6	0	1	1	-99	6	10	3	0	3
x_7	1	0	1	360	5	3	6	1	3
x_8	0	0	0	6000000	8	3	7	0	2
x_9	0	1	1	1500000	5	3	2	0	2
x_{10}	1	0	2	-99	6	11	3	0	4

对上述数据进行处理，得到离散化的矩阵为：

表 3.2　离散化后的数据

对象	*D*	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
x_1	0	3	3	1	1	3	2	0	4
x_2	1	3	2	1	1	4	2	0	3
x_3	1	1	4	1	1	2	2	0	3
x_4	1	0	1	1	1	2	2	0	4
x_5	1	2	0	1	2	3	1	0	4
x_6	0	1	1	1	2	3	1	0	3
x_7	1	0	1	2	1	1	3	1	3
x_8	0	0	0	3	3	1	4	0	2
x_9	0	1	1	2	1	1	2	0	2
x_{10}	1	0	4	1	2	4	1	0	4

在得到离散化的数据之后，首先要构造如表 3.3 所示的决策辨识矩阵，进行属性约简。

表 3.3　决策辨识矩阵

O*/*A	x_1	x_2	。。。	x_9	x_{10}
x_1	∅	*beh*	。。。	∅	*abdef*
x_2	*beh*	∅	。。。	*abceh*	∅
x_3	*abeh*	∅	。。。	*bceh*	∅
x_4	*abe*	∅	。。。	*aceh*	∅
x_5	*abdf*	∅	。。。	*abcdefh*	∅
x_6	∅	*abdef*	。。。	∅	*abeh*
x_7	*abcefgh*	∅	。。。	*acfgh*	∅
x_8	∅	*abcdefgh*	。。。	∅	*bcdefh*
x_9	∅	*abceh*	。。。	∅	*abcdefh*
x_{10}	*abdef*	∅	。。。	*abcdefh*	∅

根据决策辨识矩阵得到数据核心属性为 a、b、c、d、e，冗余属性为 f、g、h，因此对属性 f、g、h 进行约简。

其实从上述属性的含义可以看出，f 为攻击类型，而不同的攻击类型对应的死亡总数（a），受伤总数（b），财产损失的价值（c）也不同，可以看成 a、b、c 三个属性与攻击类型 f 冗余。g 为持续天数，与决策 D 有关系，可关系不大，因此，在上述小样本示例中被约掉的可能性就很大。h 为财产损害程度，本就与财产损失的价值（c）近似，被约去是降低重复。

条件属性以及决策对论域的分类分别为：

$$\boldsymbol{U}/ind(\boldsymbol{A})=\{1,2,3,4,5,6,7,8,9,10\}$$
$$\boldsymbol{U}/ind(\boldsymbol{d})=\{(1,6,8,9),(2,3,4,5,7,10)\}$$

依次去掉属性 a，b，c，d，f 得到新的分类为：

$$\boldsymbol{U}/ind(\boldsymbol{A}-a)=\{1,2,3,4,(5,6),(7,9),8,10\}$$
$$\boldsymbol{U}/ind(\boldsymbol{A}-b)=\{(1,2),(3,4),(5,6),7,8,9,10\}$$
$$\boldsymbol{U}/ind(\boldsymbol{A}-c)=\{1,2,3,(4,7),5,6,8,9,10\}$$
$$\boldsymbol{U}/ind(\boldsymbol{A}-d)=\{1,2,3,(4,6),5,7,8,9,10\}$$
$$\boldsymbol{U}/ind(\boldsymbol{A}-e)=\{1,2,3,(4,7),5,6,8,9,10\}$$

由此可得 $\boldsymbol{A}$ 的 $\boldsymbol{d}$ 正域为：

$$POS_{\boldsymbol{A}}(\boldsymbol{d})=\{1,2,3,4,5,6,7,8,9,10\}$$

同理：

$$POS_{\boldsymbol{A}-a}(\boldsymbol{d})=\{1,2,3,4,8,10\}$$
$$POS_{\boldsymbol{A}-b}(\boldsymbol{d})=\{7,8,9,10\}$$
$$POS_{\boldsymbol{A}-c}(\boldsymbol{d})=\{1,2,3,5,6,8,9,10\}$$
$$POS_{\boldsymbol{A}-d}(\boldsymbol{d})=\{1,2,3,5,7,8,9,10\}$$
$$POS_{\boldsymbol{A}-e}(\boldsymbol{d})=\{1,2,3,5,6,8,9,10\}$$

由上述结果可以得到$|POS_{\boldsymbol{A}-a}(\boldsymbol{d})|$=6，且$|\boldsymbol{U}|$为 10，$|POS_{\boldsymbol{A}}(\boldsymbol{d})|$=10，根据公式（3.10）可得：

$$I_{\boldsymbol{A}}(\boldsymbol{d})=1, I_{\boldsymbol{A}-a}(\boldsymbol{d})=0.6$$

同理可得：

$$I_{\boldsymbol{A}-b}(\boldsymbol{d})=0.4, I_{\boldsymbol{A}-c}(\boldsymbol{d})=0.8$$
$$I_{\boldsymbol{A}-b}(\boldsymbol{d})=0.8, I_{\boldsymbol{A}-c}(\boldsymbol{d})=0.8$$

利用公式（3.12）可得：

$$SGF_{\boldsymbol{A}-a}=I_{\boldsymbol{A}}(\boldsymbol{d})-I_{\boldsymbol{A}-a}(\boldsymbol{d})=0.4$$
$$SGF_{\boldsymbol{A}-b}=I_{\boldsymbol{A}}(\boldsymbol{d})-I_{\boldsymbol{A}-b}(\boldsymbol{d})=0.6$$
$$SGF_{\boldsymbol{A}-c}=I_{\boldsymbol{A}}(\boldsymbol{d})-I_{\boldsymbol{A}-c}(\boldsymbol{d})=0.2$$
$$SGF_{\boldsymbol{A}-d}=I_{\boldsymbol{A}}(\boldsymbol{d})-I_{\boldsymbol{A}-d}(\boldsymbol{d})=0.2$$
$$SGF_{\boldsymbol{A}-e}=I_{\boldsymbol{A}}(\boldsymbol{d})-I_{\boldsymbol{A}-e}(\boldsymbol{d})=0.2$$

最后结合公式（3.13）得到各个属性的权重：

$$\omega(a)=\frac{SGF_{\boldsymbol{A}-a}}{SGF_{\boldsymbol{A}-a}+SGF_{\boldsymbol{A}-b}+SGF_{\boldsymbol{A}-c}+SGF_{\boldsymbol{A}-d}+SGF_{\boldsymbol{A}-e}}=0.25$$
$$\omega(b)=\frac{SGF_{\boldsymbol{A}-b}}{SGF_{\boldsymbol{A}-a}+SGF_{\boldsymbol{A}-b}+SGF_{\boldsymbol{A}-c}+SGF_{\boldsymbol{A}-d}+SGF_{\boldsymbol{A}-e}}=0.375$$
$$\omega(c)=\frac{SGF_{\boldsymbol{A}-c}}{SGF_{\boldsymbol{A}-a}+SGF_{\boldsymbol{A}-b}+SGF_{\boldsymbol{A}-c}+SGF_{\boldsymbol{A}-d}+SGF_{\boldsymbol{A}-e}}=0.125$$

$$\omega(d)=\frac{SGF_{A-d}}{SGF_{A-a}+SGF_{A-b}+SGF_{A-c}+SGF_{A-d}+SGF_{A-e}}=0.125$$

$$\omega(e)=\frac{SGF_{A-e}}{SGF_{A-a}+SGF_{A-b}+SGF_{A-c}+SGF_{A-d}+SGF_{A-e}}=0.125$$

综上所述，可以得到 a、b、c、d、e 属性的支持度、重要性与权重，结果见表 3.4。

表 3.4　属性的支持度、重要性和权重

指标属性	支持度	重要性	权重
a	0.8	0.4	0.25
b	0.7	0.6	0.375
c	0.8	0.2	0.125
d	0.8	0.2	0.125
e	0.8	0.2	0.125

得到权重系数基础之上，结合图 3.3 的处理流程，将表 3.1 中的数据构建公式（3.14）所表述的目标属性决策矩阵，利用公式（3.15）、（3.16）进行规范化处理得到评价矩阵，再结合公式（3.18）将权重系数与规范化处理后的结果相乘，得到加权标准化的决策 $\boldsymbol{C}$ 矩阵为：

表 3.5　各个属性的 $\boldsymbol{C}$ 矩阵

对象	a	b	c	d	f
x_1	0.75	1.125	0.125	0.125	0.375
x_2	0.75	0.75	0.125	0.125	0.5
x_3	0.25	1.5	0.125	0.125	0.25
x_4	0	0.375	0.125	0.125	0.25
x_5	0.5	0	0.125	0.25	0.375
x_6	0.25	0.375	0.125	0.25	0.375
x_7	0	0.375	0.25	0.125	0.125
x_8	0	0	0.375	0.375	0.125
x_9	0.25	0.375	0.375	0.125	0.125
x_{10}	0	1.5	0.125	0.25	0.5

由公式（3.19）、（3.20）计算得到加权规范化矩阵 $\boldsymbol{C}$ 的正、负理想解分别为：

$$\boldsymbol{C}^{+}=[0.75,1.5,0.375,0.375,0.5]$$

$$\boldsymbol{C}^{-}=[0,0,0.125,0.125,0.125]$$

利用公式（3.21）、（3.22）、（3.23）得到如表 3.6 所示的正、负理想解的距离 S_i^{+}、S_i^{-} 与各个目标相对于 $\boldsymbol{C}^{+}$的相对贴近度 CL_i 以及分类情况。

表 3.6　模型预测结果

对象	正理想解	负理想解	贴近度	分类
x_1	1.965926	2.426686	0.635441	4
x_2	1.866025	2.344423	0.556811	3
x_3	2.207107	2.078298	0.484971	3
x_4	3.426686	0.965926	0.219898	2
x_5	2.974874	2.565926	0.4631	3
x_6	2.931852	1.56066	0.347391	2
x_7	3.392611	0.965926	0.221617	2
x_8	2.703143	1	0.270041	2
x_9	2.880139	1.612372	0.358902	2
x_{10}	1.719579	0.790671	0.3150	2

梳理事件的梳理与描述。

（4 级）x_1：201701010001：一名袭击者向在土耳其伊斯坦布尔奥塔科伊附近的雷纳夜总会外庆祝新年的平民开火，至少有 39 人遇难。

该事件相较于其他事件而言，屠杀平民且死亡人数过多，将其定性为 4 级事件，但从度量值上看，只是刚好超过四级事件的门限值，并不是很重要的四级事件。

（3 级）x_2：201702220014：袭击者袭击了尼日尔蒂拉贝里蒂洛阿附近的一支军事巡逻队。在袭击中至少有 15 名士兵丧生，19 名士兵受伤。

该事件针对军方的有计划的恐怖袭击，伤亡人数，尤其是军方的伤亡人数较多。虽然量化为 3 级事件，可其度量值已经接近四级事件。

（3 级）x_3：201703300002：袭击者向巴基斯坦俾路支省 Pirandar 的一个巴基斯坦武装部队车队开火。在袭击中至少有一名士兵被打死，另外三名士兵受伤。没有一个组织声称对此次袭击负责。

该事件死亡造成军方人员伤亡，明显是有预谋的恐怖事件，针对军方的事件，如果针对平民，结果会极其恶劣，因此定性为三级事件是可以接受的。

（2 级）x_4：201704070046：袭击者向巴基斯坦俾路支省 Kech 区的 Frontier Corps（FC）车队开火。袭击中至少有两名士兵受伤。

该事件没有造成人员伤亡，更没有造成财产的大量损失，定性为二级事件，同时根据度量值，刚好超过二级事件门限。

（3 级）x_5：201705250037：袭击者，包括一名自杀式炸弹袭击者，袭击了 Wahat 的叙利亚武装部队（苏丹武装部队）士兵，叙利亚霍姆斯的 Sawannah。事件中至少有 9 人死亡，其中包括 8 名士兵和炸弹袭击者。

该事件造成一定量的人数伤亡，性质恶劣，定性为三级事件。

后续的事件并没有造成大量的人员伤亡与财产损失，均定性为二级事件，不进行深入分析，罗列出具体事件与分级。

（2 级）x_6：201706010006：在伊拉克 Al Anbar 的 Al-Suwaib 引爆的一辆爆炸装置。爆炸中至少有一名伊拉克志愿军士兵被打死，一名平民受伤。这是 Hit 在同一天发生的两次类似攻击之一。没有任何组织声称对这些事件负责。

（2 级）x_7：201707110063：五名袭击者进入一个家，并在巴拉圭康塞普西翁的 Horqueta 绑架了一名女子 Maria Liduvina Ramirez Gimenez。受害者是巴拉圭人民军（EPP）的一名着名成员的嫂子，在第二天被释放之前遭到了殴打。

（2 级）x_8：201708280027：袭击者点燃了智利洛斯里奥斯地区的 29 辆测井车。袭击没有人员伤亡。

（2 级）x_9：201709060009：袭击者在哥伦比亚塞萨尔市的 Diego Hernandez 的一辆装甲车上开火。在这次袭击中，一名雇员死亡，另一名雇员受伤；此外，袭击者还从车上偷走了 150 万美元，然后放火焚烧。没有任何组织声称对这起事件负责；然而，消息来源认为这次袭击是哥伦比亚民族解放军（ELN）所为。

（2 级）x_{10}：201710240028：在马里莫普提 Tenenkou，一辆地雷引爆了一辆军用车辆。两名士兵在爆炸中受伤。JAMAAT-NASRAT al-IsAn Wal-Mulimin（JNIM）声称对这起事件负责。

从上述的分析与研究来看，本组构建的方法完全基于数据驱动，中间没有认为参与，最终用事实与数据说话。上述 10 个事件的量化评估结果与历史史实接近，可见本文方法具有良好的数据处理能力。

3.3.2 子问题 1（B）求解

由于上一节已经选取 10 个事件，并完成对其评估与分析。为体现不完备信息系统条件下的时间评估，本节依旧选取上一节中 2017 年的 10 个事件构建已知方案集，并选取 8 个形成条件属性，建立如表 3.7 所示的原始样本决策数据，得到样本集 $\boldsymbol{U}=\{x_1,x_2,\ldots,x_{10}\}$，指标（属性）集 $\boldsymbol{A}=\{a,b,c,d,e,f,g,h\}$，决策属性 D。

其中，x_1：201701010001；x_2：201702220014；x_3：201703300002；

x_4：201704070046；x_5：201705250037；x_6：201706010006；

x_7：201707110063；x_8：201708280027；x_9：201709060009；

x_{10}：201710240028。a：死亡总数（nkill）；b：受伤总数（nwound）；

c：财产损失的价值（美元）（propvalue）；

d：武器类型（weapontype1）；

e：地区（region）；f：攻击类型（attacktype1）；

g：持续天数（extended）；

h：财产损害程度（propextent）。

D：疑似恐怖主义（doubtterr）。

表 3.7　原始数据

对象	D	a	b	c	d	e	f	G	h
x_1	0	39	69	-99	5	10	2	0	4
x_2	1	15	19	-99	5	11	2	0	*
x_3	1	1	3	-99	5	6	2	0	3
x_4	1	0	2	-99	5	6	2	0	4
x_5	1	9	0	-99	6	10	3	0	4
x_6	0	1	1	-99	6	10	3	0	3
x_7	1	0	1	360	5	3	6	1	3
x_8	0	0	0	6000000	8	3	7	0	2
x_9	0	1	1	1500000	5	3	2	*	2
x_{10}	1	0	2	-99	6	11	*	0	4

对上述数据随机隐去，得到结果如下表所示。

表 3.8　不完备的原始数据

对象	*D*	*a*	*b*	*c*	*d*	*e*	*f*	*G*	*h*
x_1	0	39	69	-99	5	10	2	*	4
x_2	1	15	19	-99	5	11	2	0	*
x_3	1	1	*	-99	*	6	2	0	3
x_4	1	0	2	*	5	6	2	*	4
x_5	1	*	0	-99	6	10	3	0	4
x_6	0	1	1	-99	*	10	3	0	3
x_7	1	0	1	360	5	*	6	1	3
x_8	0	0	*	6000000	8	3	7	0	*
x_9	0	1	1	1500000	5	3	2	*	2
x_{10}	1	0	2	-99	6	11	*	0	4

隐去即缺失数据用“*”表示。对上述矩阵进行离散化，得到矩阵为：

表 3.9　离散化后的数据

对象	*D*	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
x_1	0	3	3	1	1	3	2	*	4
x_2	1	3	2	1	1	4	2	0	*
x_3	1	1	*	1	*	2	2	0	3
x_4	1	0	1	*	1	2	2	0	4
x_5	1	*	0	1	2	3	1	0	4
x_6	0	1	1	1	2	3	1	0	3
x_7	1	0	1	2	1	*	3	1	3
x_8	0	0	*	3	3	1	4	0	2
x_9	0	1	1	2	1	1	2	*	2
x_{10}	1	0	4	1	2	4	*	0	4

在得到离散化的数据之后，首先要构造如表 3.10 所示的决策辨识矩阵，进行属性约简。

表 3.10　决策辨识矩阵

O*/*A	x_1	x_2	。。。	x_9	x_{10}
x_1	∅	*be*	。。。	∅	*abde*
x_2	*be*	∅	。。。	*abceh*	∅
x_3	*aeh*	∅	。。。	*bceh*	∅
x_4	*abe*	∅	。。。	*aceh*	∅
x_5	*bdf*	∅	。。。	*abcdefh*	∅
x_6	∅	*abdef*	。。。	∅	*abeh*
x_7	*abcfh*	∅	。。。	*acfh*	∅
x_8	∅	*abcdefg*	。。。	∅	*bcdeh*
x_9	∅	*abce*	。。。	∅	*abcdeh*
x_{10}	*abde*	∅	。。。	*abcdefh*	∅

根据决策辨识矩阵得到数据核心属性为 *a*、*b*、*c*、*e*，冗余属性为 *d*、*f*、*g*、*h*，因此对属性 *d*、*f*、*g*、*h* 进行约简。

其他过程与之前的 RS 处理相似，计算决策约简集 ***B*** 中各属性的权重，通过计算可

得：

$I(B)$=0.778

$\mathrm{Sig}_B(b)$= I(B)- I($B\backslash\{a\}$)=0.778-0.75=0.028

$\mathrm{Sig}_B(b)$= I(B)- I($B\backslash\{b\}$)=0.778-0.694=0.084

$\mathrm{Sig}_B(b)$= I(B)- I($B\backslash\{c\}$)=0.778-0.528=0.25

$\mathrm{Sig}_B(b)$= I(B)- I($B\backslash\{e\}$)=0.778-0.694=0.084

在加权阈值容差关系下，取阈值 ω=0.4 时，对象 x_4 不为“*”的属性权重和为 0.376＜ω，不可能再与其他任何对象同属于一个类中，独自属于一个加权阈值容差类，将 x_4 从论域 U 中排除。剩余各对象的加权阈值容差类分别为：

$$[x_1]_B^{WT(0.4)}=\{x_1\} \qquad [x_2]_B^{WT(0.4)}=\{x_2,x_6,x_7,x_9\}$$

$$[x_3]_B^{WT(0.4)}=\{x_3\} \qquad [x_5]_B^{WT(0.4)}=[x_8]_B^{WT(0.4)}=\{x_5,x_8\}$$

$$[x_6]_B^{WT(0.4)}=\{x_2,x_6\} \qquad [x_9]_B^{WT(0.4)}=\{x_2,x_7,x_9\}$$

可以看出，加权阈值容差关系不仅考虑了属性的权重，而且可以将不满足加权阈值条件的对象预先排除，在不影响类的完整性的前提下，使得类的判定更为合理。

再根据（3.40）、（3.41）式可得加权阈值容差关系的上、下近似为：

$$B^{WT(\omega)}(X_1)=\{x_1\} \qquad B_{WT(\omega)}(X_1)=\{x_1\}$$

$$B^{WT(\omega)}(X_2)=\{x_2,x_3,x_6,x_7,x_9\} \qquad B_{WT(\omega)}(X_2)=\{x_3\}$$

$$B^{WT(\omega)}(X_3)=\{x_2,x_7,x_9\} \qquad B_{WT(\omega)}(X_3)=\varnothing$$

$$B^{WT(\omega)}(X_4)=\{x_5,x_8\} \qquad B_{WT(\omega)}(X_4)=\{x_5,x_8\}$$

进而得到属性权重为：

$\omega(a)$=0.363； $\omega(b)$= 0.261； $\omega(c)$= 0.188； $\omega(e)$=0.188

得到权重系数基础之上，结合图 3.4 的处理流程，缺失的数据利用中位数补全，将权重系数与规范化处理后的结果相乘，得到加权标准化的决策 $\boldsymbol{C}$ 矩阵为：

表 3.11 各个属性的 $\boldsymbol{C}$ 矩阵

对象	a	b	c	f
x_1	1.0905	0.783	0.188	0.564
x_2	1.0905	0.522	0.188	0.752
x_3	0.3635	1.044	0.188	0.376
x_4	0	0.261	0.188	0.376
x_5	0.727	0	0.188	0.564
x_6	0.3635	0.261	0.188	0.564
x_7	0	0.261	0.376	0.188
x_8	0	0	0.564	0.188
x_9	0.3635	0.261	0.564	0.188
x_{10}	0	1.044	0.188	0.752

由公式（3.19）、（3.20）计算得到加权规范化矩阵 $\boldsymbol{C}$ 的正、负理想解分别为：

$$\boldsymbol{C}^+=[1.0905,1.044,0.564,0.752]$$

$$\boldsymbol{C}^-=[0,0,0.188,0.188]$$

利用公式（3.21）、（3.22）、（3.23）得到如表 5.6 所示的正、负理想解的距离 S_i^+、S_i^- 与各个目标相对于 $\boldsymbol{C}^+$的相对贴近度 CL_i 以及分类情况。

表 3.12　模型预测结果

对象	正理想解	负理想解	贴近度	完备信息贴近度	分类	完备信息分类
x_1	1.965926	2.542331	0.618573	0.635441	4	4
x_2	1.866025	2.017765	0.599999	0.556811	3	3
x_3	2.207107	2.058262	0.497484	0.484971	3	3
x_4	3.426686	0.944471	0.230357	0.219898	2	2
x_5	2.974874	1.465831	0.354291	0.4631	3	3
x_6	2.931852	1.72698	0.382809	0.347391	2	2
x_7	3.392611	0.944471	0.232729	0.221617	2	2
x_8	2.703143	0.613188	0.178758	0.270041	1	2
x_9	2.880139	1.72698	0.409669	0.358902	2	2
x_{10}	1.719579	0.772763	0.3180	0.3150	2	2

梳理事件的评估与描述。

相较于完备信息，非完备信息条件下毕竟缺失了部分信息，且采用 ITR-RS 算法，并利用中位数对数据补全的方法，在量化评估与分级上相差不大。且对于随机选取的事件与随机隐去数据后，且结果依旧近似，可见该方法具有一定的适用性。

同时变化比较大的是 x_2 与 x_8 两个事件，具体事件如下：

（3 级）x_2：201702220014：袭击者袭击了尼日尔蒂拉贝里蒂洛阿附近的一支军事巡逻队。在袭击中至少有 15 名士兵丧生，19 名士兵受伤。

（2 级）x_8：201708280027：袭击者点燃了智利洛斯里奥斯地区的 29 辆测井车。袭击没有人员伤亡。

x_2 事件中不完备信息系统中评级接近于 4 级，从事件本身与完备信息下的评估来看，其事件已经接近于四级恐怖事件，虽然定性有偏差，可定量评估参数接近。

x_8 事件中不完备信息系统中评级为 1 级，由于没有造成人员伤亡，且经济损失附件 1 中没有定量给出，地区也不是敏感地区，攻击类型也只是普通的放火，因而评为 1 级也是可以接受的。同时其评价值接近于 2 级的门限，也是符合实际评估的可能。

从上述的分析与研究来看，本组构建的方法完全基于数据驱动，中间没有认为参与，最终用事实与数据说话。上述 10 个事件的量化评估结果与历史史实接近，可见本文方法具有良好的数据处理能力。

3.3.3 子问题 1 的结果与分析

通过上述两种模型，分别对题目中表 1 的事件，放在当年的数据中进行评估，得到结果如下：

表 3.13 典型事件危害级别

事件编号	危害级别	危害评价值
200108110012	5	0.92741
200511180002	5	0.86423
200901170021	1	0.1204
201402110015	3	0.46003
201405010071	3	0.52638
201411070002	2	0.3626

201412160041	5	0.9130
201508010015	1	0.10060
201705080012	5	0.90330

上述 9 个事件的描述为：

（5 级）200108110012：安哥拉完全独立全国联盟（安盟）埋设的一枚地雷使载有安哥拉罗安达和多多之间的难民的火车脱轨。武装分子用反坦克雷拦住火车，然后向火车车厢开火。大约 259 人在袭击中丧生或死亡，超过 160 人受伤。

该事件造成大量的平民死亡，危害程度为 0.92741，性质很恶劣。

（5 级）200511180002：两名身份不明的自杀式炸弹袭击者在伊拉克 Khanaqin 对两座什叶派清真寺发动了协同袭击。轰炸机进入清真寺，炸死至少 75 人，炸伤 90 人。没有一个组织声称对袭击负责。

该事件造成大量的平民死亡，危害程度为 0.86423，性质很恶劣。对比可以看出，同样是 5 级事件，其危害程度也有所区别，本文方法可以更加细化量化出同等级别的恐怖事件的结果。

（1 级）200901170021：这起事件发生在刚果民主共和国东方 Tora 的住宅区，伤亡人数不确定。该事件伤亡人数不确定，数据，尤其是重点缺失较多，评估为 1 级。

（3 级）201402110015：在伊拉克尼尼微省摩苏尔区的 Aynal-Jahish 村，一名为输油管道提供安全设施的军营被击落。至少 15 名士兵在袭击中丧生，其中八人被短暂绑架并斩首。

（3 级）201405010071：袭击者袭击了中非共和国 Kemo 县的 Mala 镇。至少 30 人在这次袭击中丧生。没有一个组织声称对这起事件负责，然而，据说归咎于 Seleka。

这个事件与上一个事件虽然都是三级事件，但量化评估明显高于上一个事件，结果更有说服力。

后续事件只是罗列事件史实，不再论述。

（2 级）201411070002：一个路边炸弹引爆附近的客轮在 Chinari 村，联邦管理部落地区，巴基斯坦。爆炸造成五人死亡，至少两人受伤。这是同一天发生在 Chinari 村的两起相关爆炸事件之一；第二次袭击发生在第一批反应人员从第一起爆炸中撤离伤员的时候。Jamaat-ul-Ahrar 声称对这些事件负责，并称他们进行袭击是为了报复安全部队据称对穆斯林社区成员的压迫性待遇。

（5 级）201412160041：袭击者在伊拉克安巴尔省 Fallujah 市杀害了 150 名妇女。没有组织声称对这起事件负责；然而，消息来源将袭击归咎于伊拉克伊斯兰国和黎凡特（ISIL），指出受害者拒绝参加圣战婚姻。

（1 级）201508010015：袭击者袭击了土耳其奥斯曼尼耶省的警察人员。一名警官在袭击中丧生，另一名受伤。没有一个组织声称有责任，但消息人士把这一事件归咎于库尔德工人党（PKK）。

（5 级）201705080012：袭击者袭击了阿富汗昆都士 Nawabad 昆都士塔哈尔公路上一批未知的警察哨所。在这场持续到 2017 年 5 月 10 日的冲突中，至少 50 人死亡，包括 10 名警官和 40 名袭击者，另有 20 名袭击者受伤。塔利班声称对这起事件负责。

3.3.4 子问题 2 的求解与结果分析

根据上述两个方法，对过去 20 年的恐怖袭击事件进行量化评估，得到每一年的结果，并得到每一年的危害均值，利用公式（3.46）与（3.47）计算，并排序得到前十位的事件如下表所示。

表 3.14 十大事件的相对危害值列表

序号	事件编号	危害相对评价值
1	200109110004	1.879384954
2	200409010002	1.871118796
3	201406150063	1.870965508
4	201710140002	1.870113448
5	200403210001	1.838242462
6	201607020002	1.824409695
7	199808070002	1.814579219
8	201404150089	1.809890553
9	200811260006	1.809802674
10	200403110004	1.808868758

上述事件在当年的评估体系中，量化的结果均接近于 1，采用相对值刻画，可以将 20 年的结果拉平，有助于数据的公平比对。

上述十个事件的详细论述如下：

1. 200109110004：“9.11”事件

这是美国发生的四起相关袭击之一，这些袭击被统称为 9/11 恐怖袭击。在当地时间上午 8:46 发生的第一次袭击中，美国航空公司 11 号航班在纽约市世贸中心大楼北塔坠毁。五名属于基地组织的劫机者在一架从马萨诸塞州波士顿洛根国际机场飞往洛杉矶国际机场的航班上控制了波音 767 飞机。机上共有 76 名乘客、11 名机组人员和五名劫机者，全部遇难。南塔在当地时间上午 9 点 59 分坍塌后，当地时间上午 10:28 在北塔倒塌。至少有 2767 人死于纽约市的袭击事件。超过 16000 人受伤。包括本·拉登和哈立德·谢赫·穆罕默德在内的基地组织领导人在多次录像采访中声称对袭击事件负责。

2. 200409010002：“别斯兰人质事件”

一群三十至三十五名武装的车臣和印古什叛乱分子，包括男人和女人，其中许多人系着自杀式炸弹腰带，在俄罗斯北奥塞梯普拉沃贝雷泽尼区的别斯兰占领了一所学校。肇事者在学校体育馆里劫持了大约 1200 名儿童、家长和教师人质。根据围困的结论，727 人受伤，大约 344 人丧生。

3. 201406150063：

攻击者绑架了大约 1686 名士兵，来自伊拉克萨拉丁省提克里特市的营地警察。两名俘虏逃脱了监管，至少 1570 人，如果不是所有其余的受害者，都被推定死亡。伊拉克伊斯兰国和黎凡特（ISIL）声称对此负责，并声称这些袭击是对杀害 ISIL 领导人 Abdul-Rahman al-Beilawy 的报复。

4. 201710140002：

索马里摩加迪沙霍丹街区 K5 十字路口，一名自杀炸弹手在 Safari 酒店外引爆了一辆载有炸药的卡车。除了袭击者之外，至少 588 人在爆炸中丧生，包括 3 名美国公民，316 人受伤。这是摩加迪沙两起自杀式袭击事件之一，袭击者打算袭击亚丁·阿德国际机场。青年党声称对这些事件负责。

5. 200403210001

尼泊尔共产党（毛派）成员袭击了尼泊尔 Bedi 的一个小镇。在袭击中，毛派分子轰炸了一座桥，抢劫了一家国有银行，轰炸了当地行政办公室，从当地监狱释放了囚犯，轰炸了一座机场。政府军与毛派进行了长期战斗，据报道，毛派杀害了 500 名毛派分子和 18 名安全人员，200 名毛派分子和 16 名安全人员受伤。

6. 201607020002

一名自杀式爆炸者在伊拉克巴格达 Karada 附近的一个购物中心引爆了一辆装有爆炸物的汽车。除了轰炸机，爆炸中至少有 382 人死亡，200 人受伤。伊拉克伊斯兰国和黎凡特（IsIL）声称对此次袭击负责。

7. 199808070002

自杀式袭击者在肯尼亚内罗毕的美国大使馆外引爆了一辆汽车炸弹，造成 224 人死亡，其中包括 12 美国人。四千人在袭击中受伤，这是基地组织犯下的。

8. 201404150089

在南苏丹联合州本图镇，袭击者袭击了一座用作民用避难所的清真寺，并绑架了一些人。这次袭击造成至少 287 人死亡，400 人受伤。这是 2014 年 4 月 15 日本提乌镇发生的五起袭击事件之一。没有组织声称对这起事件负责；然而，消息来源认为这次袭击是苏丹反对派人民解放运动发动的。

9. 200811260006

星期三晚上 2300 点左右，四名武装袭击者对泰姬陵和塔之栈酒店进行了大规模袭击。这个地点是孟买各个地点八次协同攻击的最后一次，三天内共造成 171 人死亡，250 人受伤。泰姬陵宫和塔楼饭店的四名袭击者被确认为哈菲兹·阿沙德·别名巴达·阿卜杜勒·雷曼、贾维德·别名阿布·阿里、肖艾布·别名索赫布和纳粹·别名阿布·乌默。整个晚上，他们使用手榴弹、自动武器和爆炸物在酒店造成火灾。到 0530 点，火势已经被控制住了，但是袭击者在酒店的俱乐部里劫持了 100-150 名人质。警察、军队和消防救援人员负责逮捕袭击者，释放人质，控制火势。11/29/2008 1400 时，官员重新控制泰姬陵，所有袭击者都被杀。德干圣战组织是一个先前未知的组织，声称对这次袭击负责。索赔尚未得到证实。官员们怀疑巴基斯坦的虔诚军。

10. 200403110004

当地时间早上 7 点 42 分，一枚炸弹在西班牙马德里经过圣尤金尼亚车站的一列火车的第四节车厢内爆炸。至少有 17 人丧生，更多人受伤。炸弹被伊斯兰极端分子放在背包里，用手机引爆。这次袭击是一系列十枚炸弹的一部分，这些炸弹在马德里上下班高峰期的火车上爆炸，造成 191 人死亡，1800 多人受伤。

4 问题二的建模与求解

4.1 问题分析

本问题可以拆分为以下几个子问题：一是确定 2015、2016 年度发生的、尚未有组织或个人宣称负责的 12367 起恐怖袭击事件负责人是谁，从而将 Unknown 变为 known，然后，然后采用问题 1 中的量化评估模型得到 2015 年与 2016 年事件的量化结果，将同一个组织或个人的危害度量值累加，确定排在前 5 的组织或个人，记为 1-5 号；二是建立表 2 列出的恐怖袭击事件与上述五个嫌疑人嫌疑程度的映射关系，并将结果填入表 2 中。具体解决流程如图 4.1 所示：

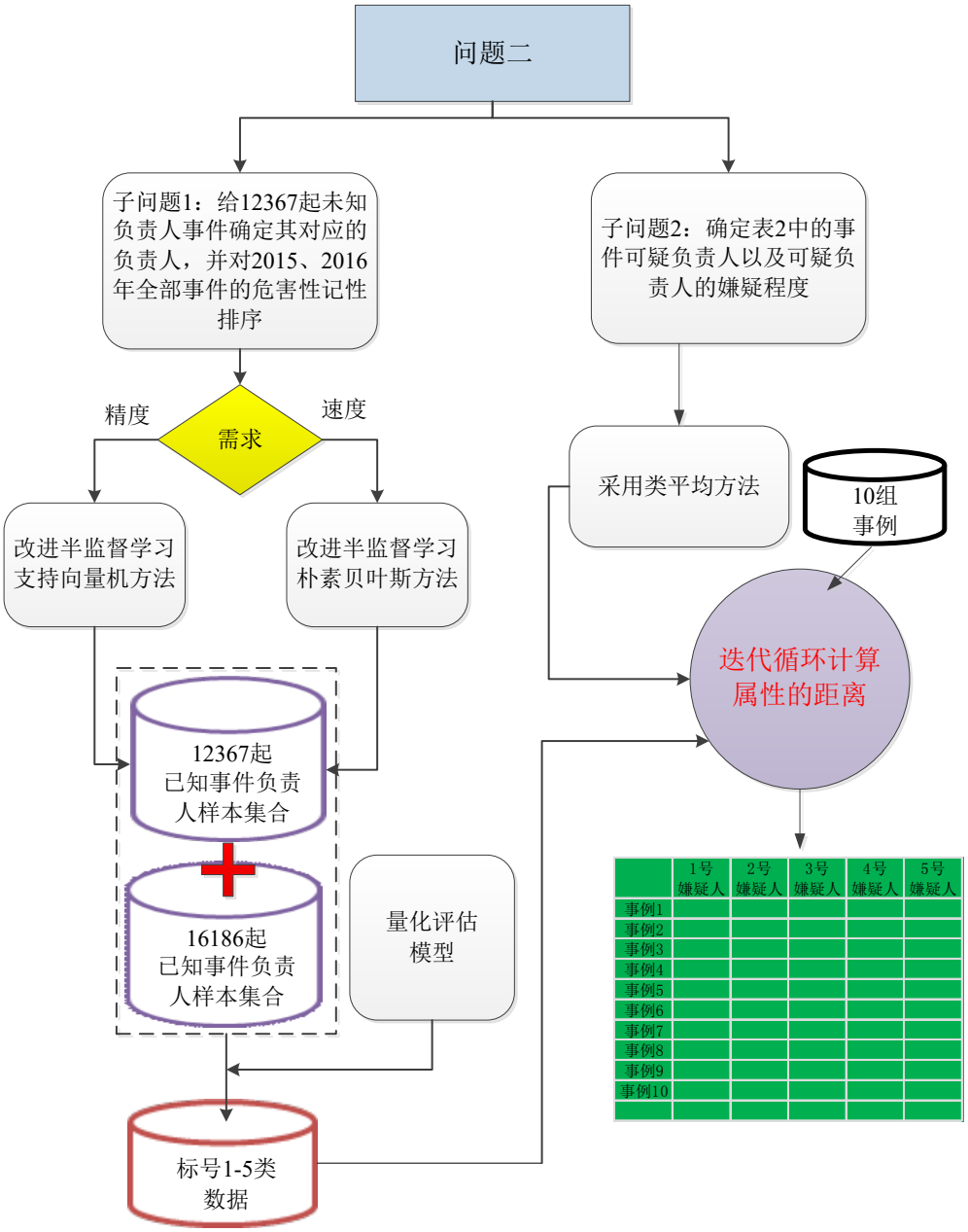


图 4.1 问题二求解思路

针对子问题 1：附件 1 中，2015、2016 年度共发生 28553 起恐怖袭击事件，其中，有组织或个人宣称负责的恐怖袭击事件有 16186 起，记为带标签数据样本，而尚未有组织或个人宣称负责的恐怖袭击事件有 12367 起，记为无标签数据。

要给这 12367 起事件带上标签，常用的方法是机器学习方法，按照任务类型主要分为：无监督学习、监督学习和半监督学习。无监督学习的目标是发现输入数据集中潜藏的结构或者规律，不适用于给未知类别标签的数据确定类别标签；另外，监督学习的目标是从已标记训练样本中学习得到样本特征和样本类别的分类器，该方法需假设具有足够多的已标记样本，从宏观上来说，已有标签数据量大于未知标签数据量，但是从具体类别中却并不满足该假设，比如在 16186 起已知标签数据中，类别为 Abbala extremists 仅有 4 个样本，类别为 Abdul Ghani Kikli Militia 仅有 1 个样本，可用作训练的样本过少，因此该方法也不适用于本问题求解。

而半监督学习的原则是通过大量无标记样本数据辅助少量已标记样本数据进行学习，可以将无标签数据依次作为输入集，样本数据标签作为输出集，从而实现无标签数据集到有标签数据的转变。为此，本问题拟采用半监督学习方法进行求解。而支持向量机（SVM）模型具有全局最优，非线性，解的稀疏性，推广能力强等优势。为此，本团队基于半监督学习，提出一种改进的半监督支持向量机模型。同时，考虑到本问题还是一个多分类问题，且样本数据相对较大，故在构造 S3VM 多分类器时采用间接法。由于半监督学习过程时间开销较大，所以本团队针对实际情况对模型运行速度的需求不同，设计了改进的半监督朴素贝叶斯模型，该模型可提高运算速度。从而实现对 12367 起事件确定标签，之后采用问题一中已建立的量化评估模型，得到 15,16 两年事件的危害量化结果，从而得到子问题 1 的解。

针对子问题 2：题中要求不仅要确定表 2 中事件对应的嫌疑人，而且还要衡量其嫌疑程度。为此，本组采用两种方法进行研究。第一种是利用五个组织的数据，构建 SVM 模型进行训练，从而将题目中要求的事件输入网络中进行结果预判，但由于该模型每次只能输出一组可能的嫌疑人，本团队将每一个事件重复输入该网络 100 次，根据输出的嫌疑人的频次排序，从而获得嫌疑度的排序。另一种方法是根据数据的相似性，基于数据之前的相似程度，采用类平均聚类的方法，将 10 组事件，逐个与将五组嫌疑人的所涉及的所有事件进行聚类，根据聚类的聚类树形图的数据关联度，来确定各个组织或个人对每个事件的可能的嫌疑程度。相比较而言，第二种方法更为直观与快捷，可精度没有第一种更为准确。根据实际情况，当要明确确定第一嫌疑人时，建议采用第一种方法，只运行一次直接得到第一嫌疑人。而当要找出可能的多个嫌疑人时，建议采用第二种方法。

4.2 模型建立

4.2.1 子问题 1 建模

4.2.1.1 子问题 1（A）模型

（1）支持向量机模型

支持向量机（Support Vector Machines, SVM）是基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。给定训练样本集 D=$\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，$y_i \in \{-1,+1\}$，分类学习的基本思想就是基于训练集

D 在样本空间找到一个划分超平面，将不同类别的样本分开。

如图 4.2 所示，距离超平面最近的几个训练样本点满足上式，被称为“支持向量”。训练完成后，大部分训练样本都不需保留，最终模型只与支持向量有关，说明支持向量的选取对 SVM 的学习训练具有不容忽视的作用，这也正是“支持向量机”名字的由来。

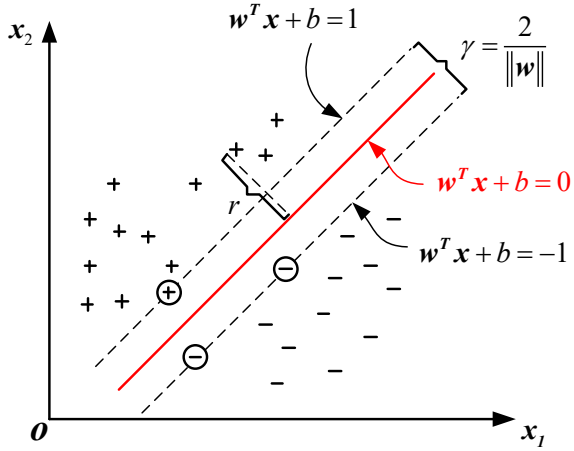


图 4.2 支持向量机分类效果示意图

为了获得最大间隔的划分超平面，可设计如下模型：

$$\begin{aligned} &\min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ &s.t. \ \ y_i\left(w^T x_i + b\right) \geq 1-\xi_i, \ \ i=1,2,\cdots m \qquad \xi_i \geq 0, i=1,2,\cdots m \end{aligned} \tag{4.1}$$

在目标函数中引入惩罚系数 C，其作用是调节学习算法在特征空间中的置信范围与经验风险的比例，从而优化学习算法的泛化能力。引入拉格朗日乘子 $\alpha_i \geq 0$，将上式转化为求其对偶问题，并利用“核函数”技巧，则有：

$$\begin{aligned} &\max_{\alpha}\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ &s.t.\sum_{i=1}^{m}\alpha_i y_j = 0, \qquad 0 \leq \alpha_i \leq C, \qquad i=1,2,\cdots,m \end{aligned} \tag{4.2}$$

支持向量机算法继集成了最大间隔超平面、凸二次优化、稀疏解和松弛变量等技术，根据有限的样本能获得很好的学习能力，但也存在一些局限：当样本规模较大时训练过程收敛速度明显放慢，时间复杂度较高；核函数的选取缺乏理论依据等。针对以上问题，学者们提出多种支持向量机改进算法，具有代表性的有：序列最小优化算法、SVM Light 算法、变种-边界约束支持向量机等。

（2）半监督 SVM

半监督支持向量机（Semi-supervised Support Vector Machines, S3VM）是传统支持向量机在半监督学习上的推广，主要思想是从训练学习产生的多个大边缘低密度分类器中寻求最优分类器的过程。S3VM 算法现已在文本分类、图像检索、生物信息学、语言处理等各个领域应用广泛。在不考虑未标记样本时，支持向量机视图找到最大间隔划分超平面，而在考虑未标记样本后，S3VM 试图找到能将两类有标记样本分开，且穿过数据低密度区域的划分超平面，如图 4.3 所示（“+”“-”分别表示有标记的正反例，灰色点表示未标记样本）。这里的“低密度分隔”（low-density separation）,显然这是聚类假设在考虑了线性超平面划分后的推广。

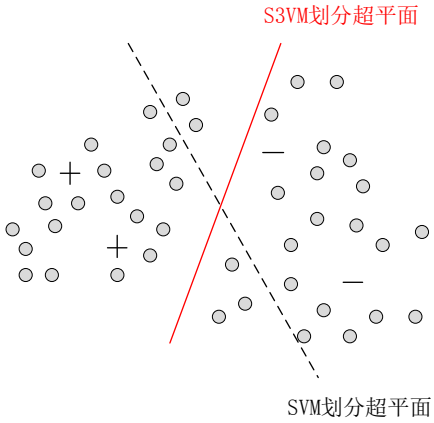


图 4.3 半监督支持向量机与低密度分隔

由图 4.3 可看出：在给定的少量有样本和大量无标记样本区域中，经过训练学习可产生多个大边缘低密度分类器，但考虑到有标记样本数量过少，导致难以判断哪个是与之最符合的最佳分类器。

半监督支持向量机中最著名的是 TSVM（Transductive Support Vector Machine）。与标准 SVM 一样，TSVM 也是针对二分类问题的学习方法，TSVM 试图考虑对未标记样本进行各种可能的标记指派。即尝试将每个未标记样本分别作为正例或反例，然后在所有这些结果中，寻求一个在所有样本（包括有标记样本和进行了标记指派的未标记样本）上间隔最大化的划分超平面。一旦划分超平面得以确定，未标记样本的最终标记指派就是其预测结果。

考虑二分类问题，给定一组有标签样本数据 $D_l=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$ 和无标签样本数据 $D_u=\{\hat{x}_{l+1},\hat{x}_{l+2},\cdots,\hat{x}_{l+u}\}$，其中 $x,\hat{x}\in\chi$，$y\in\{+1,-1\}$，l 和 u 分别是有标记样本数目和无标记样本数目（$l\ll u$）。目标是找出一个函数 $f:\chi\in\{\pm1\}$，$\hat{y}\in\{\pm1\}^u$ 解决如下的目标函数最小化问题：

$$\begin{aligned}&\min\left\{\frac{\|w\|^2}{2}+C_l\sum_{i=1}^{l}\xi_i+C_u\sum_{i=l+1}^{l+u}\xi_i\right\}\\&s.t.\quad y_i\left(w^Tx_i+b\right)\geq1-\xi_i,\ i=1,2,\cdots,l\\&\qquad\hat{y}_i\left(w^Tx_i+b\right)\geq1-\xi_i,\ i=l+1,l+2,\cdots,l+u\\&\qquad\xi_i\geq0,\ i=1,2,\cdots,l+u\end{aligned}\tag{4.3}$$

其中 (w,b) 确定了一个划分超平面；ξ 为松弛变量，$\xi_i(i=1,2,\cdots,l)$ 对应于有标记样本，$\xi_i(i=l+1,l+2,\cdots,l+u)$ 对应于未标记样本；C_l 与 C_u 是由用户指定的用于平衡复杂度、已标记样本和未标记样本的拟合误差参数。算法描述如表 4.1 所示。

表 4.1 半监督支持向量机算法

输入：有标记样本集 $D_l=\{(\boldsymbol{x_1},y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_l,y_l)\}$；
未标记样本集 $D_u=\{\boldsymbol{x}_{l+1},\boldsymbol{x}_{l+2},\ldots,\boldsymbol{x}_{l+u}\}$
折中参数 C_l，C_u.
过程：
1：用 D_l 训练一个 SVM_l

2：用SVM_l对D_u中样本进行预测，

3：初始化$C_u \ll C_l$；

4：**while** $C_u < C_l$ **do**

5:　基于D_l，D_u，$\hat{\boldsymbol{y}}$，C_l，C_u求解式(13.9)，得到$(\boldsymbol{w}, b)$，ξ；

6：　**while** $\exists\left\{i, j \middle| (\hat{y}_i \hat{y}_j < 0) \wedge (\xi_i > 0) \wedge (\xi_j > 0) \wedge (\xi_i + \xi_j > 2)\right\}$ **do**

7：　$\hat{y}_i = -\hat{y}_i$；

8：　$\hat{y}_j = -\hat{y}_j$；

9：　基于D_l，D_u，$\hat{\boldsymbol{y}}$，C_l，C_u重新求解式(13.9)，得到$(\boldsymbol{w}, b)$，ξ

10:　**end while**

11:　$C_u = \min\left\{2C_u, C_l\right\}$

12:　**end while**

输出：未标记样本的预测结果：$\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

（3）算法缺点分析

半监督支持向量机是传统支持向量机在半监督学习上的推广，主要思想是从训练学习产生的多个大边缘低密度分类器中寻求最优分类器的过程。仔细分析可发现传统的半监督支持向量机存在以下缺陷：

S3VM 算法在学习训练分类器的过程中，首先要先利用初始有标记样本数据训练出一个初始 SVM 分类器。支持向量机主要处理小样本学习问题，之所以称之为小样本机器学习方式，主要是因为支持向量机的凸二次规划求解过程与大部分的训练样本无关，而只与小部分的支持向量（Support Vector， SV）相关。式（4.2）的系数α_i是凸二次规划问题的解，其中不同的α_i对应不用的训练样本，只有当$\alpha_i \neq 0$所对应的样本才会对求解结果产生影响。这部分样本即被称为支持向量，故分类超平面只有这部分训练样本有关。但一般的 SVM 在训练过程中，使用所有的有标记样本数据进行学习求解，这样做会出现大的计算量和不必要的计算开销，同时因为多余有标记样本的加入有可能会导致学习性能的下降，进而影响识别效果。因此，在不影响 SVM 训练学习性能的情况下，若能采取某些方法提取出支持向量，而不是将所有有标记样本都参与训练学习，这样做便能减少训练样本的数目，节省运算时间。

由图 4.3 可看出，在同时存在有标记样本数据和无标记样本数据的情况下，S3VM 会训练出多个低密度大间隔分类决策面，S3VM 旨在从多个大边缘低密度分类器中找出一个最优的分类器，在此过程中，不可避免的会舍弃相当多数量的分类决策面；另外当有标记样本同时与多个分类决策面相吻合时，很难决定究竟哪一个才是最佳分类器，一旦选择错误便会影响分类准确率，从而导致学习性能的下降。

（4）基于 KCVNN 的支持向量预选取算法

由 S3VM 的原理以及支持向量的几何分布结构中可发现：SVM 训练出的分类超平面只取决于输入样本中的一小部分样本—即支持向量。为了节省计算开销，本团队采用K折交叉验证最近邻方法（K cross-validation NN algorithm，KCVNN 算法）提取出训练样本的支持向量。

该算法针对每一个有标记样本x_i，首先采用最近邻算法提取出它附近的的K个最近邻样本，然后计算这K个最近邻样本中属于相同类别的样本比例q_i，设定门限q，将满足$q + q_i < 1$部分的有标记样本选取为支持向量。参数K和q对支持向量的选取有着至关重要的影响，因此为了能够选择到最优参数，本团队采用参数K折交叉验证法确定最优

参数 K 和 q 的具体数值。

交叉验证是对分类器性能进行评价的一种方法，在样本数据量较少的情况下，通过对数据进行重复利用来评估分类的精确性。具体思想是对初始样本数据进行分组，其中一部分作为训练集，另一部分作为测试集，先采用训练集进行模型的训练，然后利用测试集对模型的准确性进行验证，如此循环 K 次，最后利用 K 次测试集精确率的平均值作为参考指标，选取那些对应于最大平均精确率的参数作为最终的最优参数。通常常用的交叉验证法由留一法交叉验证， K 折交叉验证和 $Holder-Out$ 交叉验证等。本节采用 5 折交叉验证法来寻求最优参数 K 和 q 。

图 4.4 为本团队算法的示意图，其中三角和正方形分别代表两类样本，若设定 $k=2$ 和 $q=0.7$，可以计算红色三角最近邻的两个有标记样本中，属于同类别样本的比例为 0，满足 $q+q_i=0.7<1$，故红色三角可以被选取为支持向量；类似地，对于蓝色正方形也可以被选取为支持向量。对于黄色三角而言，可以看出与它最近邻的两个有标记样本中，同类别样本的比例为 1，即满足 $q+q_i=1>0.7$，因此黄色三角样本不能被选取为支持向量。

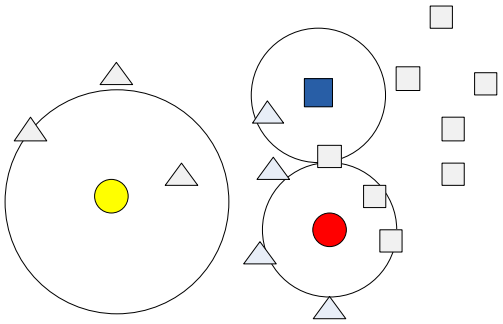


图 4.4　基于 KCVNN 算法示意图

基于 KCVNN 的支持向量预选取算法流程如下：

表 4.2 基于 KCVNN 的支持向量预选取算法

输入：有标记样本集 $D=\{(x_i,0),i=1,\cdots,n_0\}\cup\{(x_i,1),i=1,\cdots,n_1\}$；
输出：选取的支持向量集合 T；
1.将初始有标记样本集随机分成 K 份，取其中 $K-1$ 份为训练集，剩余 1 份为验证集；
2.当 $1\le K\le 5$ 时，求解 5 折交叉验证法下的精确率；
3.取 K 次测试集精确率的平均值；
3. 选取最大平均精确率对应的最优参数 K 和 q；
4.将对应的具有最高平均精确率的支持向量集合作为选取的集合 N。

（5）基于 HSS 的改进半监督支持向量机

针对 S3VM 只考虑一个最优分类器会降低分类精度的缺陷，本团队提出一种基于启发式采样搜索（Heuristic sampling search, HSS）的改进 S3VM 算法。启发式采样搜索方法的最终目的是寻求具有代表性的大边缘低密度分类决策面，具体过程分为两个阶段：首先训练出多个大边缘低密度分类器，其次采用聚类方法选出具有代表性的差异性较大的 T 个分类决策面。

用 $h(w,\hat{y})$ 表示目标函数，

$$h(w,\hat{y})=\min\{\frac{\|w\|^2}{2}+C_1\sum_{l=1}^{l}l(y_i,f(x_i))+C_2\sum_{j=1}^{u}l(\hat{y}_j,f(\hat{x}_j))\} \tag{4.4}$$

改进的 S3VM 目标是要找出多个大边缘低密度分类决策面$\{f_t\}_{t=1}^{T}$以及所对应的类划分$\{\hat{y}_t\}_{t=1}^{T}$，即要使下列函数取得最小值：

$$\min_{\{f_t,\hat{y}_t\in\beta\}_{t=1}^{T}} \sum_{t=1}^{T} h(f_t,\hat{y}_t) + M\Omega(\{\hat{y}_t\}_{t=1}^{T}) \tag{4.5}$$

其中，T 表示分类决策面的数量；Ω 是用于表示分类决策面之间差异性的惩罚函数；M 是常数，用于保证差异性。

针对上述问题不好求解的特点，引入拉格朗日乘子$\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_N\}$寻找最优分类面，将 *RBF* 作为核函数代入(4.1)，得到分类函数为：

$$f(x)=\text{sign}(\sum_i \alpha_i y_i \bullet k(x_i \bullet x)+b) \tag{4.6}$$

定义分类面距离影响因子：对于未标记样本x_i^*，定义分类面距离影响因子$S(x_i^*,w_t)$对其进行标记。若w_{t1}和w_{t2}为距离x_i^*最近的两个分类决策面，则称$S(x_i^*,w_{t1})$和$S(x_i^*,w_{t2})$为x_i^*的分类面距离影响因子。其中 w 表示训练学习后的低密度分类决策面，$S(x_i^*,w)>0$。

结合以上分析，写出改进的半监督支持向量机算法的主要步骤：

Step1 设定参数C_1、C_2，利用已有的样本训练集经过训练学习产生多个大边缘低密度分类决策面。对训练产生的这些分类决策面进行迭代计算，直到式(4.1)收敛时结束；

Step2 基于启发式抽样搜索的方法对训练集进行采样，采样出 T 个具有代表性的差异大的分类决策面，运用 K 均值聚类算法获取这 T 个分类决策面；

Step3 对于未标记样本x_i^*，采用分类面距离影响因子$S(x_i^*,w_t)$对其进行标记。若w_{t1}和w_{t2}为距离x_i^*最近的两个分类决策面，则称$S(x_i^*,w_{t1})$和$S(x_i^*,w_{t2})$为x_i^*的分类面距离影响因子。若$S(x_i^*,w_{t1})>S(x_i^*,w_{t2})$，则$x_i^*$的类别由$w_{t1}$分类决策面决定；同理，若$S(x_i^*,w_{t1})<S(x_i^*,w_{t2})$，则$x_i^*$的类别由$w_{t2}$分类决策面决定；当$S(x_i^*,w_{t1})=S(x_i^*,w_{t2})$，则采用抽签法决定$x_i^*$的类别。通过上述分析，可筛选出 T 个差异性较大的分类决策面，从而完成对未标记样本的分类。具体算法如表 4.3 给出：

表 4.3 基于 HSS 的改进半监督支持向量机算法

Input:有标记样本集$\{x_i,y_i\}_{i=1}^{l}$，未标记样本集$\{\hat{x}_j\}_{j=l+1}^{j=l+u}$，T；

Output：$\{\hat{y}_t\}_{t=1}^{T}$；

1.设定循环次数 P;

2.for $n=1:P$ do

3. while not converged do

4. fix $\{y\}$,solve $\{w,b\}$ via SVM solver;

5. optimize $\{w,b\}$ via S3VM solver;

6. end while // get separators $\{y_n\}_{n=1}^{N}$;

7.end for//get S3VM multiple low-density separators $\{\hat{y}_n\}_{n=1}^{N}$;

8.perform K -means algorithm for the representative separators where $t=T$;

9.output $\{\hat{y}_t\}_{t=1}^{T}$.

综上，基于 HSS 的改进支持向量机算法流程图如下所示：

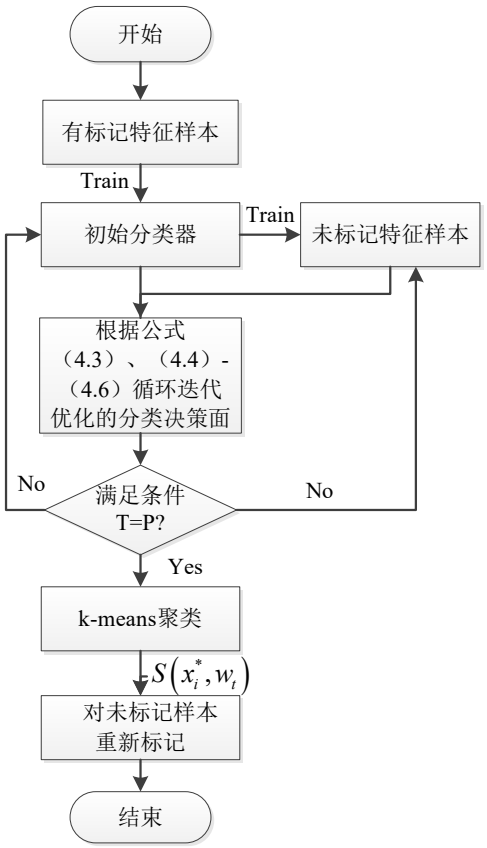


图 4.5 基于 HSS 的改进支持向量机算法流程图

4.2.1.2 子问题 1（B）模型

前文介绍了本团队提出的基于改进的 S3VM 的恐袭数据分析方法。但该算法存在一些问题：S3VM 算法的求解是一个非凸二次规划问题，该问题不具有全局最优解，求解过程需要非常高的时间复杂度，而在运算过程中运用了迭代手段，迭代次数过大，这些问题都会导致半监督支持向量机存在效率低下问题的困扰，虽然能取得较好的恐袭数据学习效果，但附件 1 中的恐怖袭击数据高达 11 万条，时间复杂度过高。本团队在贝叶斯理论的基础上，引入半监督思想，提出一种半监督朴素贝叶斯算法。贝叶斯分类器是通过最大化后验概率对恐袭数据样本属性进行判断，是机器学习中的标准分类器。朴素贝叶斯的时间复杂度很低，如果用 n 表示数据集的属性数目，用 m 表示训练集的样本数量，朴素贝叶斯的时间复杂度为 $O(n,m)$。实践证明，朴素贝叶斯分类器的性能受样本变化的影响较小，学习速度快，尤其是在满足独立性假设的情况下，朴素贝叶斯分类器在小样本训练集上也能取得很好的分类性能。但很多情况下无法确定数据集是否满足独立性假设，为减轻条件独立性假设带来的负面影响，需要对其进行改进，为此本团队引入半监督学习思想，提出改进的半监督朴素贝叶斯分类器，依次来降低对于恐袭数据分析的时间复杂度。

（1）朴素贝叶斯理论

NB 分类器的基本思想是根据贝叶斯决策论，通过待分类样本的一些特征的先验概率来计算该样本属于某分类的后验概率，从而预测该样本的类别标记。

假设恐袭训练集样本有 m 个类别分别为 $C_1, C_2, \ldots, C_m$，提取的 n 个数据特征属性分别为 $X_1, X_2, \ldots, X_n$。给定一个未知类别的数据样本 X，当且仅当 $P(C_i|X) > P(C_j|X)(1 \le i, j \le m, i \ne j)$ 时，分类器将预测 X 属于具有最高后验概率的类 $C_i(1 \le i \le m)$。可知：

$$P(C_i|X) = \frac{P(C_i)P(X|C_i)}{P(X)} \tag{4.7}$$

其中，类先验概率 $P(C_i)$ 表达了恐袭数据样本中各类样本所占的比例，根据大数定律，$P(C_i)$ 可通过各类样本出现的频率进行估计。而对类条件概率 $P(X|C_i)$ 来说，由于它涉及关于 X 所有属性的联合概率，直接根据样本出现的频率来估计将会遇到严重的困难。为了解决这个问题，NB 分类器采用了属性条件独立性假设，即：对已知类别，假设所有属性相互独立；换言之，就是假设每个属性独立地对分类结果发生影响。

基于属性条件独立性假设，式(4.7)可以重写为：

$$P(C_i|X) = \frac{P(C_i)P(X|C_i)}{P(X)} = \frac{P(C_i)}{P(X)}\prod_{k=1}^{n} P(X_k|C_i) \tag{4.8}$$

其中，X_k 为样本 X 在第 k 个属性上的取值。

由于对所有类别来说 $P(X)$ 相同，因此贝叶斯判定准则有：

$$h^*(X) = \arg\max_{1 \le i \le m} P(C_i)\prod_{k=1}^{n} P(X_k|C_i) \tag{4.9}$$

这就是 NB 分类器的表达式。显然，NB 分类器的训练过程就是基于训练集来估计类先验概率 $P(C_i)$，并为每个属性估计条件概率 $P(X_k|C_i)$。

令 s_i 表示训练集中 C_i 类恐袭数据样本数，s 表示训练集样本总数，则易估计出类先验概率：

$$P(C_i) = \frac{s_i}{s} \tag{4.10}$$

条件概率 $P(X_k|C_i)$ 可由 C_i 类训练样本信息计算产生。恐袭数据样本特征部分具有连续属性，但大部分为离散数据。对于具有连续属性的恐袭数据样本特征，假设其分布为正态分布，利用样本集特征信息对分类器进行训练学习，得到分类器的参数，即特征期望向量 $\overline{m}_i = (\overline{m}_{i1}, \overline{m}_{ij}, ..., \overline{m}_{in})$ 和方差向量 $\sigma_i = (\sigma_{i1}, \sigma_{ij}, ..., \sigma_{in})$，则根据正态分布概率公式有：

$$P(X_k|C_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{\frac{-(X_k - \overline{m}_{ij})^2}{2\sigma_{ij}^2}} \tag{4.11}$$

所以，NB 分类器连续模型可写为：

$$\begin{aligned} h^*(X) &= \arg\max_{1 \le i \le m} P(C_i)\prod_{k=1}^{n} P(X_k|C_i) \\ &= \arg\max_{1 \le i \le m, 1 \le j \le n} P(C_i)\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{\frac{-(X_k - \overline{m}_{ij})^2}{2\sigma_{ij}^2}} \end{aligned} \tag{4.12}$$

对于具有离散属性的恐袭数据样本特征，本团队假定其服从古典概型，采用频率代表概率。令某一项频数为 r_{ij}，该特征总数为 r_i，则：

$$P(X_k \mid C_i) = \frac{r_{ij}}{r_i} \tag{4.13}$$

所以，NB 分类器离散模型可写为：

$$\begin{aligned} h^*(X) &= \arg\max_{1\le i\le n} P(C_i)\prod_{k=1}^{n} P(X_k \mid C_i) \\ &= \arg\max_{1\le i\le n} P(C_i)\prod_{k=1}^{n} \frac{r_{ij}}{r_i} \end{aligned} \tag{4.14}$$

（2）改进半监督朴素贝叶斯算法

在传统的 NB 分类器中融入半监督学习的思想，可以在模型训练过程中减少大量的人力物力，同时又能使分类器获得比较好的泛化能力。但是，由半监督朴素贝叶斯（Semi-supervised Naïve Bayes, SNB）分类算法可知，如果初始 NB 分类器对测试样本的分类准确率较高，则加入到初始训练集中带有类别标记的测试样本大部分都是分类正确的，此时得到的 SNB 分类器性能会有所提升；但如果初始 NB 分类器对测试样本的分类结果准确率较低，则加入到初始训练集中带有类别标记的测试样本大部分都是分类错误的，这些错误样本在循环迭代中不仅不会使 SNB 分类器的性能提升，还可能导致分类器性能下降。针对这一问题，本文在 SNB 分类器的基础上提出了一种改进算法，以进一步提高分类性能。

本团队针对半监督朴素贝叶斯将预测的全部无标记样本添加至有标记样本集会产生迭代错误进而降低分类准确率的缺陷，提出一种改进的 SNB 算法（RSNB）。该方法通过在无标记样本集生成的置信度列表中选取置信度较高的样本，这样训练样本始终保持最优，从而更加准确地对测试样本集的类别进行预测，再利用预测后的分类结果对分类器参数（分类器参数即特征期望向量 $\overline{m}_i$ 和方差向量 σ_i ）进行改进，从而提高分类准确率以及分类实时效果。

首先在基本的 SNB 算法中，需要计算多个概率的连乘，由于很多个很小的数相乘会造成数据错误，从而会对结果造成偏差，采用对乘积取常用对数的解决方式，将连乘变成连加。同时从运算效率考虑，加法比乘法降低了分类的时间复杂度 $O(n)$。经常用对数改进后的 SNB 算法的分类器为：

$$\begin{aligned} P(C_i / X) &= \max(l+u)P(C_i) \\ &\sum_{k=1}^{n} \log P(X_k / C_i)(l+u) \end{aligned} \tag{4.15}$$

本团队提出数据置信度的概念，并对 SNB 分类器的算法进一步改进，提出一种 RSNB 分类算法。其基本原理是在用初始 NB 分类器对测试样本进行分类时，确定某一个样本类别标记的同时计算其属于该类别的分类置信度，所有样本分类结束后，将每一类测试样本按照置信度数值由高到低的顺序进行排序，然后在每一类中选取部分高置信度值的测试样本加入到初始训练集中进行第二次训练，这样可以保证有标记的训练样本集中保存的是高质量的样本。利用 RSNB 算法生成并选取高分类置信度样本的过程如图 4.6 所示，其中 Top$M$ 代表从分类置信度列表中选取的置信度较高的前 M 个样本。

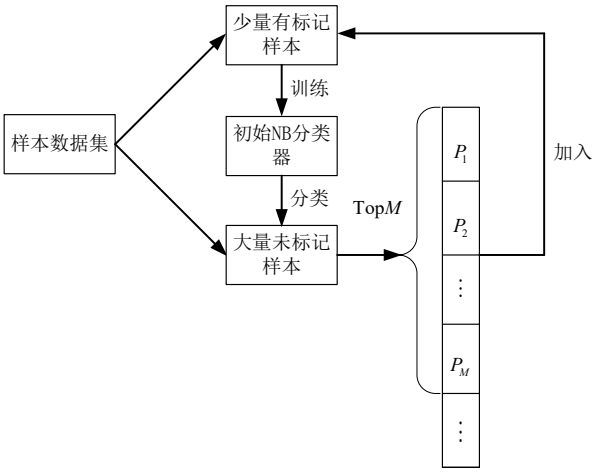


图 4.6 生成并选取高分类置信度样本过程

已知有标签样本数据集 $L:\{(x_i,y_i)\}_{i=1}^{l}$，其类别为 $C:\{C_i,\ (1\le i\le n)\}$；无标签样本数据集为 U：$\{\hat{x}_j\}_{j=1}^{u}$；数据置信度列表为 *Clist*，置信度为 A_i，选取的高置信度样本数目为 M。算法步骤为：

步骤 1：从 U 中任取样本 $\hat{x}_j$，计算属于 C_i 的后验概率 $P(C_i/\hat{x}_j)$，以及属于非 C_i 类的平均概率 $\frac{1}{N-1}\sum_{\bar{C}_i\neq C_i}P(\bar{C}_i/\hat{x}_j)$。

步骤 2：令置信度 A_i 为 $P(C_i/\hat{x}_j)$ 与 $\frac{1}{N-1}\sum_{\bar{C}_i\neq C_i}P(\bar{C}_i/\hat{x}_j)$ 之差,其中 $|C|=|C|-1$,若 $|C|$ 不为 0，则返回步骤 1。

步骤 3：返回所有 A_i 中数值最大的 A_{imax} 并将其作为 $\hat{x}_j$ 的置信度，将 A_i 添加到置信度列表 *Clist* 中，令 $|U|=|U|-1$,若 $|U|$ 不为 0，则返回步骤 1。

步骤 4：对置信度列表中的样本 $\hat{x}_j$ 按 A_i 的数值大小进行降序排列，取出前 M 个 A_i 所对应的样本 $\hat{x}_j$，将其添加至有标号样本集 L 中，从而更新样本集，完成对分类器的改进过程。

从以上流程可以看出，RSNB 分类器与 SNB 分类器的主要区别在于新增测试样本的选择方法。在 SNB 分类器中，新增数据是在测试样本中随机选择的；而在 RSNB 分类器中，新增数据是测试样本中分类置信度数值较高的数据，这样就提升了训练集数据的质量，可以使更多分类正确的测试样本加入到初始训练集中，使训练样本始终保持最优，再利用预测后的分类结果对分类器参数（即特征期望向量 $\overline{\boldsymbol{m}}$ 和方差向量 $\boldsymbol{\delta}$）进行修正，从而进一步提高分类准确率。

另外，算法中选取的高置信度样本数 M 越小，即加入到初始训练集 D 中的数据量越小，RSNB 算法越接近于 NB 算法；反之 M 越大，即加入到初始训练集 D 中的数据量越大，则 RSNB 算法越接近于 SNB 算法。在实际应用中，M 值并不是固定不变的，而是应该根据初始 NB 分类器对测试集的测试准确率确定。如果初始 NB 分类器对测试集的测试准确率比较高，则应该选择较大的 M 值，使得测试集中更多分类正确的样本加入到初始训练集中，以提高分类器的性能；如果初始 NB 分类器对测试集的测试准确率比较低，则应该选择较小的 M 值，避免测试集中过多分类错误的样本加入到初始训练集中导致分类器的性能降低。

基于 RSNB 算法的恐袭数据分类流程主要分为四个阶段，具体如图 4.7 所示。

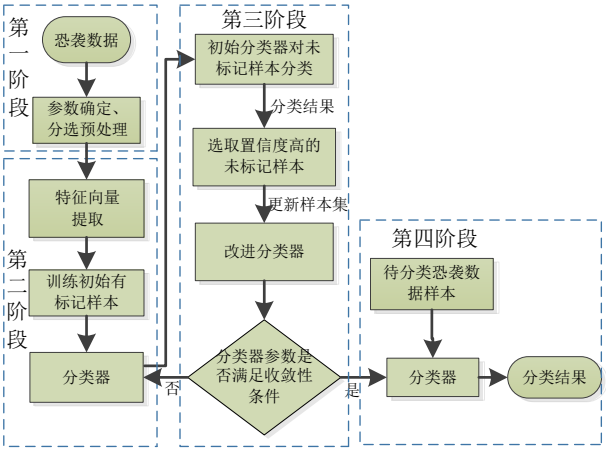


图 4.7 基于 RSNB 算法的恐袭数据分类流程

第一阶段为恐袭数据预处理阶段。将待训练的恐袭数据进行参数确定，分选预处理。

第二阶段为恐袭数据初始有标记特征样本的选取阶段，具体为：

（1）提取恐袭数据的样本特征 $X=(x_{i_1},x_{i_2},...,x_{i_M})$,选取出 M 个有标记样本作为初始训练样本，利用初始训练样本的特征 X 对分类器进行学习，计算得出恐袭数据初始分类器参数，即均值向量 $\overline{m}_i^0=(\overline{m}_{i1}^0,\overline{m}_{i2}^0,...,\overline{m}_{in}^0)$ 和方差向量 $\sigma_i^0=(\sigma_{i1}^0,\sigma_{i2}^0,...,\sigma_{in}^0)$。

（2）若 $i=m$，则输出分类器参数，即特征期望向量 $\overline{m}^0=(\overline{m}_1^0,\overline{m}_2^0,...,\overline{m}_m^0)$ 和方差向量 $\sigma^0=(\sigma_1^0,\sigma_2^0,...,\sigma_n^0)$。否则 i=i+1，返回（1）。

第三阶段为恐袭数据未标记特征样本的选取阶段，具体为：

（1）利用初始分类器对恐袭数据的未标记特征样本进行分类。

（2）在每次更新迭代过程中，选择未标记特征样本集中预测准确程度比较高的 M 个样本，将其加入至有标记特征样本集中，使用更新过的特征样本集对分类器进行改进，输出改进后的分类器参数特征期望向量 $\overline{m}^{i+1}=(\overline{m}_1^{i+1},\overline{m}_2^{i+1},...,\overline{m}_m^{i+1})$ 和方差向量 $\sigma^{i+1}=(\sigma_1^{i+1},\sigma_2^{i+1},...,\sigma_n^{i+1})$ 向量。

（3）判断分类器参数是否满足收敛性条件（其中 α 为可调节参数）$\left\|\overline{m}^{i+1}-\overline{m}^{i}\right\|+\left\|\sigma^{i+1}-\sigma^{i}\right\|\le\alpha$，若满足则输出最终分类参数特征期望向量 $\overline{m}=(\overline{m}_1^{i+1},\overline{m}_2^{i+1},...,\overline{m}_m^{i+1})$ 和方差向量 $\sigma^{i+1}=(\sigma_1^{i+1},\sigma_2^{i+1},...,\sigma_n^{i+1})$；若不满足，则返回（1）。

第四阶段为对恐袭数据的分类阶段。分别提取出待分类恐袭数据的特征向量作为输入，利用改进后的分类器对测试样本（无标记特征样本）的类别进行预测，分类过程完成。

4.2.2 子问题 2 建模

4.2.2.1 子问题 2（A）模型

子问题 2 不仅要确定表 2 中事件对应的嫌疑人，而且还要衡量其嫌疑程度。另外，此问题中，对于精度的需求要求高，为此，将上述子问题 1 中半监督 SVM 模型简化为经典的 SVM 模型，即，全监督网络进行计算。具体计算流程如下所示：

Step1：为保证数据样本的足够丰富，现从附件 1 中抽取犯罪嫌疑人 1-5 的所有恐怖袭击事件作为初始数据样本。

Step2：将初始数据随机划分成 10 组互不相交的子集，其中的 9 组作为训练数据，1 组作为测试数据，输入到子问题 1 中的 SVM 模型中。设定当训练精度达到 90%，表明此半监督 SVM 学习网络已训练好。

Step3：依次将表 2 中的事例作为输入集，用训练好的 SVM 模型来确定其嫌疑人，并记录输出结果。

Step4：重复 Step3 100 次，并记录所有的输出结果

Step5：对输出结果进行统计分析，计算各个嫌疑人的出现频率，按照频率由大到小的顺序排列，其中，嫌疑人的嫌疑程度与嫌疑人出现的频率成正相关，并将结果依次填入表 2 中。

需要指出的是：采用此种方法，可能出现运行 100 次得到的嫌疑人个数不足 5 个，且即使你的运行次数足够多，也可能会出现犯罪嫌疑人的个数不足 5 个，但是，题中指出：“如果认为嫌疑人关系不大，也可以保留空格”，所以采用本问题的方法是符合题意的。

4.2.2.2 子问题 2（B）模型

为进一步完善子问题 2 中（A）模型的不足，本团队采用类平均法的思想解决该问题，具体实现步骤如下：

Step1：在子问题 1 的基础上，将筛选出的前 5 个事件负责人视为 5 名嫌疑犯；

Step2：从附件 1 中抽取出 5 名嫌疑犯所有事例，并将所有事例的属性按照问题 1 中模型进行离散化，组成数据集 X。X 被定义为：

$$X=\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \tag{4.16}$$

其中，p 是恐怖事例的属性维度，矩阵的每一行对应一个事例 x_i，而 x_{if} 表示事例 x_i 的第 f 个属性值。需要指出的是，所有事例中的属性已将负责人这一属性剔除。

Step3：将表 2 中的事例也按照 Step2 的方法进行离散化，并表示为式（4.16）的形式；

$$X'=\begin{bmatrix} x'_{11} & \cdots & x'_{1f} & \cdots & x'_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_{i1} & \cdots & x'_{if} & \cdots & x'_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_{101} & \cdots & x'_{10f} & \cdots & x'_{10p} \end{bmatrix} \tag{4.17}$$

Step4：计算 $X'(i,:)$ 与 $X(j,:)$ 之间的距离欧式距离 d_{ij}：

$$d_{ij}=\sqrt{\sum_{f=1}^{p}(x'_{if}-x_{jf})^2} \tag{4.18}$$

Step5：记录所有计算结果，得到 D：

$$D=\begin{bmatrix} d_{11} & \cdots & x_{1q} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ip} & \cdots & x_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{101} & \cdots & x_{10p} & \cdots & x_{10n} \end{bmatrix} \tag{4.19}$$

Step6：对 D 按照行从小到大进行排序，取前 5 项得到 D'：

$$D'=\begin{bmatrix} d'_{1i_1} & d'_{1i_2} & d'_{1i_3} & d'_{1i_4} & d'_{1i_5} \\ d'_{2j_1} & d'_{2j_2} & d'_{2j_3} & d'_{2j_4} & d'_{2j_5} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d'_{10k_1} & d'_{10k_2} & d'_{10k_3} & d'_{10k_4} & d'_{10k_5} \end{bmatrix} \tag{4.20}$$

Step7：找到对应事件的负责人，从而得到子问题 2 的解。

4.3 模型解算及结果分析

4.3.1 子问题 1（A）求解及结果分析

通过前文论述，由于附件一中存在一定量的组织，其制造的恐怖袭击事件并不多，因此，采用全监督学习的方式并不能够取得良好的效果。而半监督学习可以利用少量样本对数据进行学习，并能够很好的利用无标签数据，从而进一步提升了数据的识别度。

为了说明本文所提出的算法相较于主流的全监督学习与经典的半监督学习，具有更好的效果。本节首先选择一定量已知组织者的恐怖组织的事件，即有标签的事件。之后对这些数据按一定的比例去标签，构建少部分已知标签，而大部分没有标签的新的数据集。将本文的方法与其他的全监督学习与半监督学习方法，对构建的新的数据集进行学习，之后对那些去标签的数据重新标签化。将得到的结果与原始数据集进行比对，根据识别率量化本文方法的优势。

为量化本文算法的处理结果，首先选定 2015 年到 2016 年 Al-Shabaab、Boko Haram、Islamic State of Iraq and the Levant (ISIL)、Taliban 四个组织所做的事件的参数。

对于每一个组织事件，在范围内随机选取样本集的 5%、10%、15%、20%、25%作为已标记样本，剩余的作为未标记样本（测试样本），对传统半监督支持向量机算法(S3VM)和改进后的 S3VM 算法进行了对比，所选取的核函数类型和参数都相同。实验中，S3VM 核函数采用高斯径向基核函数 RBF。实验结果如表 4.4 所示。（改进的 S3VM 算法用 Improved S3VM 表示）

表 4.4 两种方法在不同标记样本下的识别效果比较

有标识样本的比例	5%		10%		15%		20%		25%	
不同算法	S3VM	IS3VM	S3VM	IS3VM	S3VM	IS3VM	S3VM	IS3VM	S3VM	IS3VM
Al-Shabaab	77.32	77.10	78.35	77.89	77.98	81.36	80.93	83.51	79.83	81.87
Boko Haram	76.41	77.64	76.92	79.56	78.65	81.29	80.05	83.21	80.50	82.44
ISIL	75.90	77.48	76.90	79.25	78.51	81.97	80.55	83.07	79.90	82.88

Taliban	76.14	77.86	76.05	79.91	78.78	81.01	80.39	82.44	79.79	83.30
平均识别率/%	76.14	77.86	76.05	79.91	78.78	81.01	80.39	82.44	79.79	83.30

（1）识别率

从表一中实验数据结果可以看出：（1）在标记样本分别占5%、10%、15%、20%、25%的样本中，改进后的S3VM算法识别率明显高于传统的S3VM算法，这说明改进后的算法具有明显的优势。改进后的S3VM算法综合考虑具有代表性的多个差异性较大的大边缘低密度分类决策面，避免了S3VM只选取一个分类决策面会影响分类精度的缺陷，在分类准确率方面具有很大的提高。下面比较下不同规模的初始有标记样本集数据在两种算法下对平均分类性能的影响，实验结果如图4.9所示。

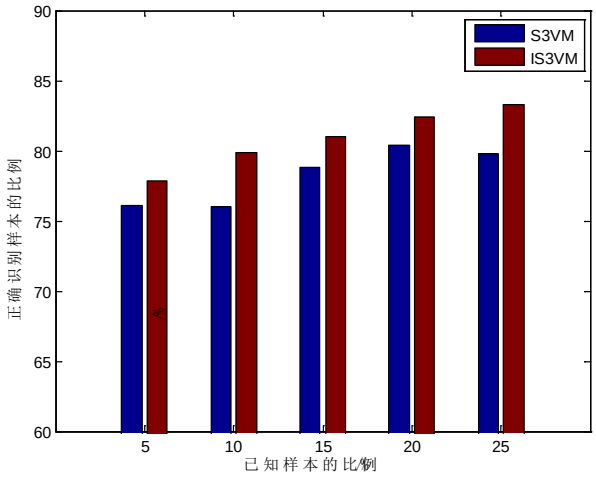


图 4.9 两种算法识别率比较

图 4.9 中，横轴为有标记样本集规模，纵轴为识别率。从图 4.9 中可看出，本文提出的改进 S3VM 算法的识别率随着传统的 S3VM 算法识别率的提高而升高，改进的 S3VM 的识别率始终高于传统 S3VM 算法的识别率。在改进的 S3VM 算法中，当有标记样本比例达到 20%时识别率最高，这表明充足的训练样本集（20%）构造的具有代表性的分类决策面差异性最大，从而分类性能达到最佳。

（2）不同分类器实时性分析

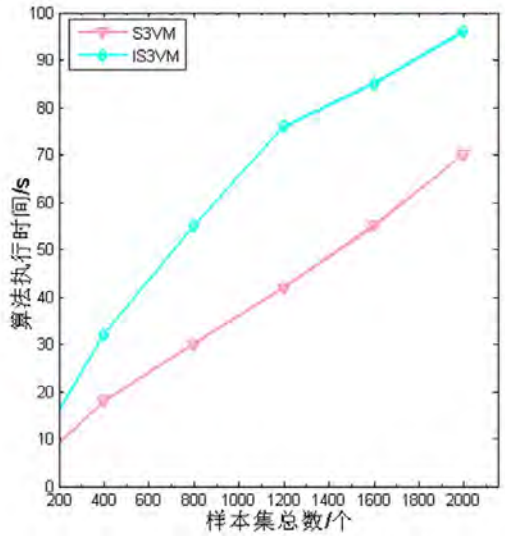


图 4.10 两种算法实时性分析

随着有标记样本比例的逐渐增大，T_{s3vm}、T_{Is3vm}也随之增加，由于IS3VM考虑到多个分类决策面，所以在时间消耗上有T_{Is3vm}＞T_{s3vm}。这主要是因为S3VM算法的求解是一个非凸二次规划问题，求解过程需要非常高的时间复杂度，而在运算过程中运用了迭代手段，迭代次数过大，这些问题都会导致半监督支持向量机存在效率低下问题的困扰，虽然IS3VM算法在时间消耗上多于S3VM算法，但结合识别率来看，改进算法识别率明显高于原算法，说明改进算法在整体的效率方面具有很大的提升。

（3）不同算法识别效果比较

为验证本文算法的有效性，在有标记样本比例为20%时，将本文提出的算法与 S3VM 算法， BP 神经网络法、SVM 进行对比。对比结果如表 4.5 所示。

表 4.5　不同算法识别率对比

已知标签比例/%	5	10	15	20	25
S3VM	76.14	77.05	78.78	80.39	79.79
IS3VM	77.86	79.91	81.01	82.44	83.30
BP	55.47	60.03	65.53	68.03	71.72
SVM	58.74	62.69	67.25	70.43	73.27

上图中，横轴为已知标签比例，纵轴为三种方法在不同识别比例下的识别率。从结果中可看出，改进后的 S3VM 算法在不同比例时识别率明显大于其余两种算法。在比例比较低时，导致 BP 神经网络对识别的分类性能明显降低，但此时 IS3VM 算法仍然能够保持较高的识别率，说明 IS3VM 算法在提高分类性能上具有很大的优势。

因此，本组提出的 IS3VM 方法无论是速度，精度还是效果均强于现有的方法，更适合于半监督学习。

通过上述仿真验证了本组所提出的方法的效能，且根据本组成员之前的技术积累，该方法已经在其他领域内取得良好效果。因此，用于确定恐怖事件的嫌疑人具有更好的效果。

4.3.2 子问题 1（B）求解及结果分析

为量化本文算法的处理结果，依旧选定 2015 年到 2016 年 Al-Shabaab、Boko Haram、Islamic State of Iraq and the Levant (ISIL)、Taliban 四个组织所做的事件的参数。

对于每一个组织的事件，在范围内随机选取样本集的 5%、10%、15%、20%、25%作为已标记样本，剩下的作为未标记样本（测试样本），将改进前后算法进行了对比，实验结果如下：

表 4.6　两种方法在不同标记样本下的识别效果比较

有标识样本的比例	5%		10%		15%		20%		25%	
不同算法	SNB	RSNB	SNB	RSNB	SNB	RSNB	SNB	RSNB	SNB	RSNB
Al-Shabaab	66.04	67.37	66.93	69.35	68.26	70.83	69.24	72.52	67.88	72.52
Boko Haram	65.35	66.80	68.82	68.47	68.18	71.19	68.81	71.70	68.21	73.23
ISIL	65.96	67.28	66.92	69.36	68.20	71.17	69.44	71.96	67.32	72.40
Taliban	66.34	67.62	67.33	68.96	68.68	66.60	69.06	71.78	68.56	71.81
平均识别率/%	65.93	67.27	67.51	69.04	68.34	70.74	69.14	72.78	68.03	72.49

（1）识别率

对于每一个组织，随着有标记样本比例的增加，改进后的 SNB 算法识别率明显高于传统的 SNB 算法，这说明改进后的算法具有明显的优势。但也有部分特征样本集中 RSNB 算法识别率低于 SNB 算法，这是因为当组织事件样本集中大多数都是准确率较高的预测结果时，SNB 的方法能够使得更多的数据加入到有标记样本集中，而采用本文方法时，添加至训练样本集的已知事件集有限，从而使得更新后的样本集中可靠的样本数据量减少，因此识别率略低于 SNB 算法。两种算法在对四种组织的样本特征集上的平均识别率如图 4.11 所示。

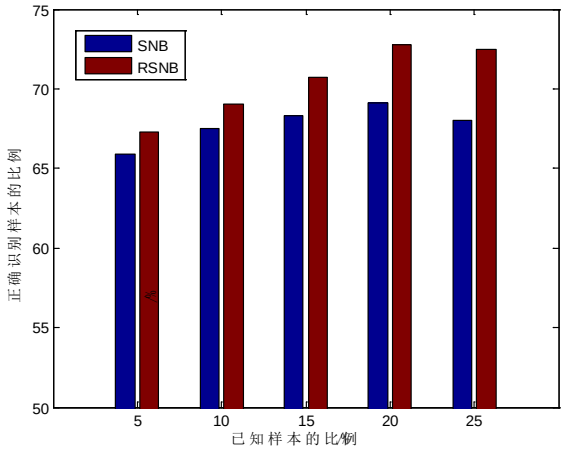


图 4.11 两种算法平均识别率对比

从图 4.11 可看出，RSNB 算法较 SNB 算法相比具有更好的分类性能。当有标记样本比例过小或过大时，识别率都会有所下降，从表中可看出有标记样本比例达到 20%时识别率最高，此时分类效果较理想且所需的标记数据量也较少。该方法通过选取高置信度的无标记样本添加到训练集中，使得训练样本集中始终保留的是高质量的样本，从而避免了传统的 SNB 将预测的全部无标记样本添加至有标记样本集会产生迭代错误的缺陷，提高了对测试样本集的分类准确率。

（2）分类器实时性分析

为进一步对比两种算法构建的分类器的实时性，本文以 FMCW 信号为例，在有标记样本比例为 20%时，分别计算两种算法在不同数量的有标记样本条件下的运行时间，实验结果如图 5.8 所示。

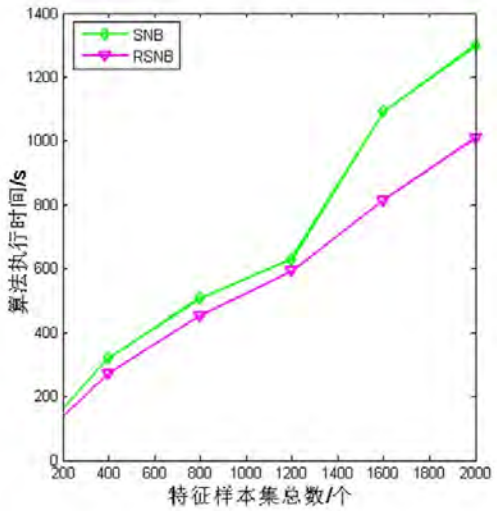


图 5.8 两种算法执行时间对比

实验结果表明：当提取的信号特征样本数据量持续增加时，RSNB 算法相比于传统的 SNB 算法耗费的时间更小，且数据集越大算法执行时间越小。因为 RSNB 算法舍弃了一部分预测准确低的样本，使得训练集总数减少而且保留的都是高质量的样本，从而大大节省了运行时间。但是当特征样本数据量较少时，两种算法执行时间不相上下，这是因为样本数据量较少时，新算法仍需生成置信度列表，然后添加预测准确度高的一部分样本到有标记样本集中，耗费了部分时间，所以会出现在数据量较少时 RSNB 算法实时性优势不明显的情况。

（3）不同算法识别效果比较

将本文提出的 RSNB 算法与 SNB 算法、以及 PCA-SVM 算法进行对比实验，识别结果如表 4.7 所示。

表 4.7　不同算法识别率对比

已知标签比例/%	5	10	15	20	25
PCA-SVM	35.81	47.33	54.26	59.82	64.78
SNB	57.51	62.17	64.15	66.68	68.97
RSNB	58.70	62.81	68.34	71.18	73.94

RSNB 算法的识别率均高于 SNB 算法和 PCA-SVM 算法。由于 SNB 算法将预测到的所有特征样本都添加至有标记样本，会产生迭代错误进而降低识别率；而采用 PCA-SVM 算法时，对原始样本数据提取后的维数依然很高，此时训练出的 SVM 分类器学习能力不强，识别效果不太理想；RSNB 算法通过生成置信度列表进而选取预测准确度较高的特征样本，极大的提高了识别率，半监督学习的效果更有效。

通过上述仿真验证了本组所提出的方法的效能，且根据本组成员之前的技术积累，该方法已经在其他领域内取得良好效果。因此，用于确定恐怖事件的嫌疑人具有更好的效果。

通过上述两种模型可以确定未知袭击者的恐怖事件，即将 Unknown 变成 Known。再结合任务 1 构建的模型，可以确定事件的危害程度。将同一类组织或个人 15 到 16 年两年的危害性进行累加，并取出前五位，结果如下表所示。

表 4.8　危害度最高的五个组织

序号	组织名称	累计危害程度
1	Islamic State of Iraq and the Levant (ISIL)	2163.661
2	Taliban	1740.391
3	Al-Shabaab	942.679
4	Boko Haram	728.378
5	Houthi extremists (Ansar Allah)	594.979

即为任务 2 题目中要求的危害性排在前 5 位的组织或个人。

4.3.3 子问题 2（A）求解及结果分析

通过上一节，得到 1 到 5 号危害度最高的五个组织分别为：

表 4.8　危害度最高的五个组织

Islamic State of Iraq and the Levant (ISIL)
Taliban
Al-Shabaab
Boko Haram
Houthi extremists (Ansar Allah)

首先将上述五个组织的恐怖袭击事件用于训练 SVM 网络，将训练好的网络用于题目中要求的事件锦衣嫌疑人判定。将一个事件作为输入，重复 100 次，记录输出的结果的频次，近似为可能为此组织的概率。将概率由大到小排序，即可确定嫌疑人顺序。得到的结果如下所示。

表 4.8　SVM 判断的恐怖组织嫌疑度

事件编号	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	3	2	4	5	1
201705050009	1	4	5	2	3
201705050010	1	4	5	2	3
201707010028	5	4	1	2	3
201707020006	1			2	
201708110018		1			
201711010006		1			
201712010003	1				

上表中可以看出。

201701090031 事件基本上确定为 Islamic State of Iraq and the Levant (ISIL)所为。

201702210037 事件基本上确定为 Taliban 所为。

201708110018 事件基本上确定为 Taliban 所为。

201711010006 事件基本上确定为 Taliban 所为。

201712010003 事件基本上确定为 Islamic State of Iraq and the Levant (ISIL)所为。

201707020006 事件可能为 Islamic State of Iraq and the Levant (ISIL)所为，也可能为 Boko Haram 所为。

而 201703120023、201705050009、201705050010、201707010028 出现多个嫌疑人，而且从记录的频次来看，并没有像其他事件一样，几乎 100 次均指向某个组织，而是呈现近乎平均的分布。

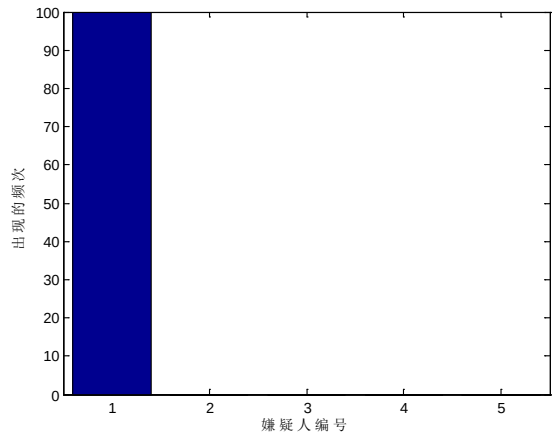


图 4.9 201701090031 事件频次分布

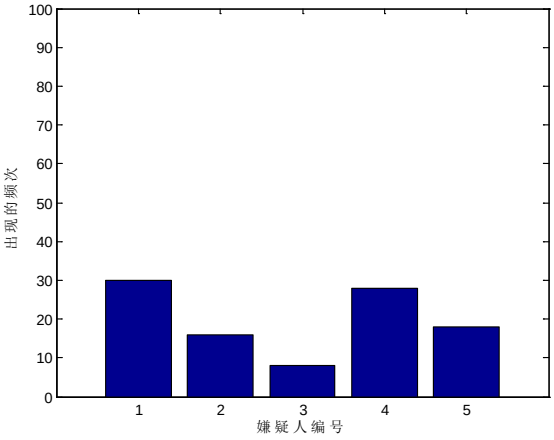


图 4.10 201705050009 事件频次分布

尤其是 201705050009 和 201705050010 两个参数几乎一致的事件，虽然排序相似，可分布区别较大。如果某个事件五个组织几乎等可能参与的，那就等价于五个组织都可能没参与。

同时，由于 SVM 自身训练的原因，只要有输入，他的输出就是五大组织中的某一个，而前提是这个事件是五大组织中的某个做的才可以，而一旦不是五大组织所为，或参数与之前的参数差距很大，将这个事件输入到 SVM 中去，依旧会得到五大组织中的某个，这样得到的结果有待进一步研究。

即，本模型适用于五大组织的确定性推断，而如果一旦出现多种可能时，其推断可能有误。

4.3.4 子问题 2（B）求解及结果分析

接下来利用类平均聚类法来描述事件，由于五大组织中各个组织事件过多，为展示清晰，将待确定事件与关系较近的事件构建聚类图，可得。

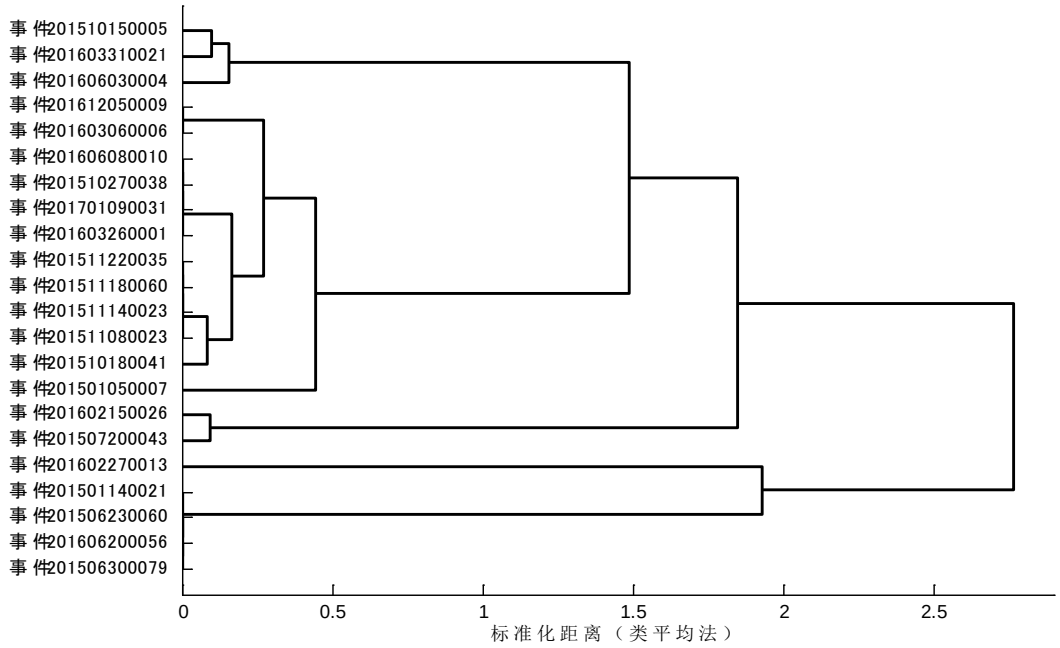


图 4.11 201701090031 与 Islamic State of Iraq and the Levant (ISIL)组织数据聚类图

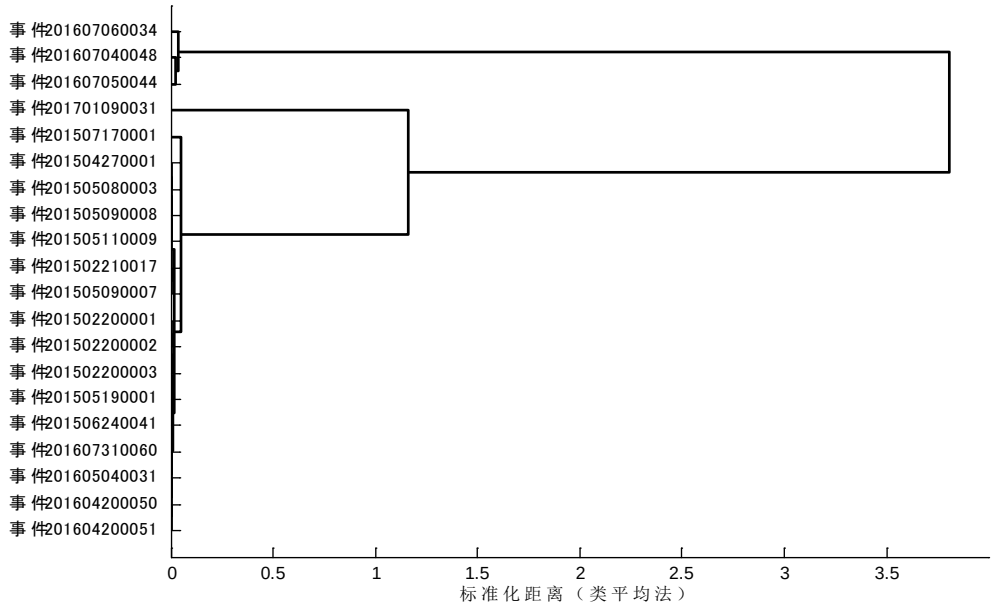


图 4.12 201701090031 与 Boko Haram 组织数据聚类图

通过两图对比可以看出 201701090031 与 Islamic State of Iraq and the Levant (ISIL)组织数据的贴近度，要明显高于 Boko Haram 组织的数据。可以更为直观的看出数据的相似性。

表 4.9　类平均聚类法判断的恐怖组织嫌疑度

	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	3	2	5	4	1
201705050009	1	4	5	2	3
201705050010	1	4	5	2	3
201707010028	5	4	1	2	3
201707020006	1			2	
201708110018		1			
201711010006		1			
201712010003	1				

上述只有一个的是因为聚类程度，即某个组织的树状图深度远小于其他的组织，因此，近似的定性为该事件为这个组织所为。

即 201701090031 事件基本上确定为 Islamic State of Iraq and the Levant (ISIL)所为。

201702210037 事件基本上确定为 Taliban 所为。

201708110018 事件基本上确定为 Taliban 所为。

201711010006 事件基本上确定为 Taliban 所为。

201712010003 事件基本上确定为 Islamic State of Iraq and the Levant (ISIL)所为。

201707020006 事件可能为 Islamic State of Iraq and the Levant (ISIL)所为，也可能为 Boko Haram 所为。

而 201703120023、201705050009、201705050010、201707010028 事件的并不大，

勉强按照聚类的结果进行排序，但并不能说明可能是这些组织所为。

上述嫌疑度排序与 SVM 结果比较接近，近似为嫌疑度的排序。

为进一步确定上述事件的可能实施者，利用本任务中的第一个将 Unknown 转换为 Known 的模型，对 2017 年的数据进行处理，得到结果如下。

表 4.10　半监督预测的结果

事件	嫌疑人
201701090031	Islamic State of Iraq and the Levant (ISIL)
201702210037	Taliban
201703120023	Sudan People's Liberation Movement in Opposition (SPLM-IO)
201705050009	Sudan People's Liberation Movement in Opposition (SPLM-IO)
201705050010	Sudan People's Liberation Movement in Opposition (SPLM-IO)
201707010028	Seleka
201707020006	Islamic State of Iraq and the Levant (ISIL)
201708110018	Taliban
201711010006	Taliban
201712010003	Islamic State of Iraq and the Levant (ISIL)

通过表 4.10 的结果可以佐证。

201701090031 事件基本上确定为 Islamic State of Iraq and the Levant (ISIL)所为。

201702210037 事件基本上确定为 Taliban 所为。

201708110018 事件基本上确定为 Taliban 所为。

201711010006 事件基本上确定为 Taliban 所为。

201712010003 事件基本上确定为 Islamic State of Iraq and the Levant (ISIL)所为。

201707020006 事件可能为 Islamic State of Iraq and the Levant (ISIL)所为。

而 201703120023、201705050009、201705050010、201707010028 事件可能并非五大组织所为。

201703120023、201705050009、201705050010 很可能为 Sudan People's Liberation Movement in Opposition (SPLM-IO 所为。

201707010028 很可能为 Seleka 所为。

5 问题三的建模与求解

5.1 问题分析

本问题主要目的是分析研判下一年全球或某些重点地区的反恐态势，提出反恐斗争的见解和建议。具体研究对象主要包括：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。

事件发生的主要原因可以理解为确定附件 1 中描述恐怖主义事件的属性重要程度，为此，可以采用粗糙集的方法来量化指标重要程度。

时空特性可以理解为建立时空特性模型，以此来确定恐怖袭击事件发生与发生的时间和发生的地点的逻辑映射关系，其研究对象侧重于某一个恐怖组织或者个人，为此，可以采用 ARMA 和 RBF 模型相结合提出一种改进的时空特性模型来解决此问题。

蔓延特性可以理解为恐怖袭击事件某个区域内，随着时间的演变，其数量的变化趋势，可借鉴传染病 SIR 模型，建立恐怖袭击事件蔓延特性模性，研究的侧重点是对恐怖袭击事件发生的趋势进行推断。

级别分布可以理解为恐怖袭击事件的危害程度在地域和时间上的分布特性。

最后，依据恐怖事件发生的主要原因、时空特性、蔓延特性以及级别分布的模型求解结果，给出反恐斗争的方案支持意见和有力举措。

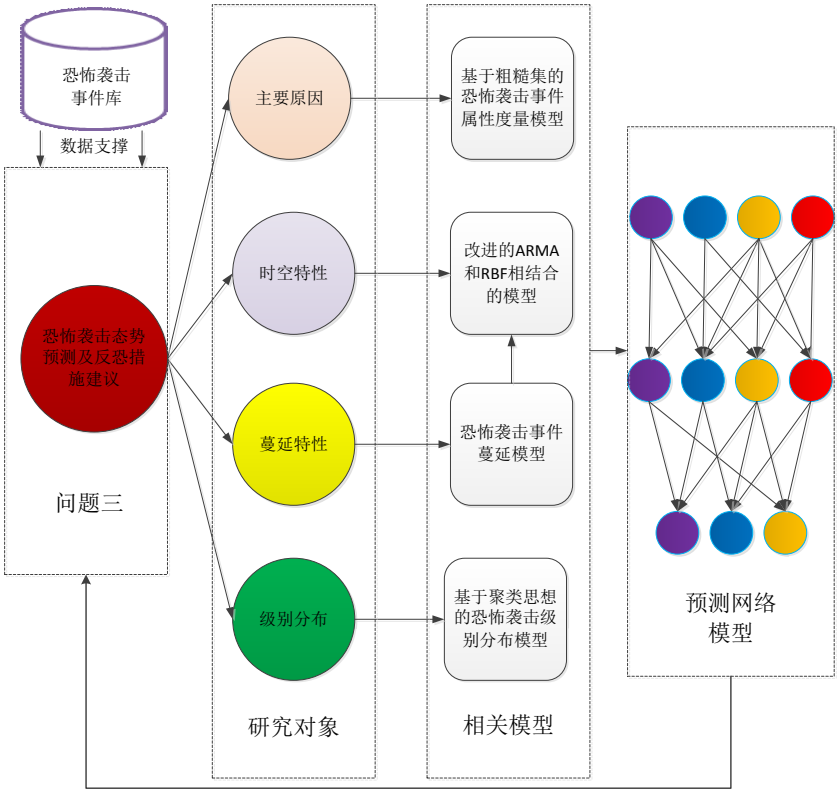


图 5.1 问题三分析

5.2 模型建立

5.2.1 基于粗糙集的恐怖袭击事件属性度量模型

为确定主要因素，根据前文的 RS-TOPSIS 或者 ITR-RS 中的方法即可确定属性权重，从而根据属性权重进行排序，即可得到主要属性。或者采用简单直接的经典粗糙集理论，同时结合根据经典粗糙集衍生出来的工具软件，Rosetta 进行处理，

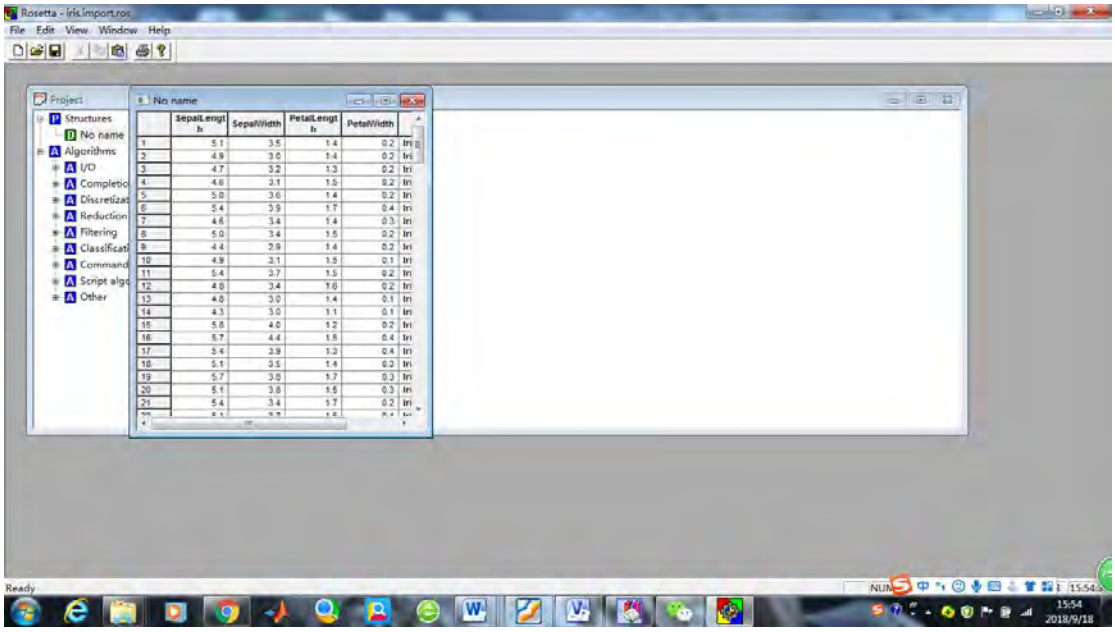


图 5.2 Rosetta 软件交互图

软件的算法原理如下。

称$\{U,A,F,d\}$是决策信息系统，其中 $U=\{x_1,x_2,\ldots, x_n\}$为对象集，$U$ 中的每个元素 $x_n(i\leq n)$称为一个对象。$A=\{a_1,a_2,\ldots, a_m\}$为属性集，$A$ 中的每个元素 a_l（$l\leq m$）称为一个属性。$F=\{f_l:U\rightarrow V_l(l\leq m)\}$为 U 与 A 之间的关系集，其中 V_l 为 a_l（$l\leq m$）的值域。$d:U\rightarrow V_d$ 为决策，V_d 取有限值。每个属性子集 $a\subseteq A$ 决定了一个不可区分的关系 $ind(A)$：

$$ind(A)=\{(x,y)\in U*U\mid \forall a\in A,a(x)=a(y)\} \tag{5.1}$$

关系 $ind(a_i)(a\subseteq A)$构成了 U 的划分，用 $U/ind(a)$表示。

设$\{U,A,F,d\}$是一个信息系统，对于任意 $B\subseteq A$，记：

$$R_B=\{(x_i,x_j)\mid f_l(x_i)=f_l(x_i)(a_l\in B)\} \tag{5.2}$$

R_B 则 U 是上的等价关系，记：

$$[x_i]_B=\{x_j\mid (x_i,x_i)\in R_B\} \tag{5.3}$$

则 $U/R_B=\{[x_i]_B|x_i\in U\}$是 U 上的划分。同理：

$$R_d=\{(x_i,x_j)\mid d(x_i)=d(x_i)\} \tag{5.4}$$

决策信息系统是将信息系统中的属性分为条件属性与决策属性两类，因此需要研究两类属性的关系，从信息系统仅仅可以得到分类，而从决策信息系统中可以获取决策知识。记：

$$R_{a_l}=\{(x_i,x_j)\mid f_l(x_i)=f_l(x_i)(a_l\in A)\} \tag{5.5}$$

属性约简流程为：

Step 1：构造决策辨识集进行属性约简；

设{*U,A,F,d*}为决策信息系统，记：

$$U/R_A=\{[x_i]_A \mid x_i \in U\} \tag{5.6}$$

$$U/R_d=\{[x_i]_d \mid x_i \in U\} \tag{5.7}$$

$$D_d([x_i]_A,[x_j]_A)=\begin{cases}\{a_l \in A \mid f_l(x_i)\neq f_l(x_j)\}, [x_i]_d \cap [x_j]_d=\varnothing \\ \varnothing \qquad [x_i]_d \cap [x_j]_d=\varnothing\end{cases} \tag{5.8}$$

称 $D_d([x_i]_A,[x_j]_A)$为$[x_i]_A$ 与$[x_j]_A$ 的决策辨识集，称：

$$D_d=(D_d([x_i]_A,[x_j]_A) \mid [x_i]_A,[x_j]_A \in U/R_A) \tag{5.9}$$

为决策信息系统的决策辨识矩阵，从而得到系统的决策辨识矩阵。

Step 2：构造决策辨识集进行属性约简；

设{*U,A,F,d*}为决策信息系统，则 *B* 为决策协调集，当且仅当对于任意的 $D_d([x_i]_A,[x_j]_A)\neq\varnothing$，有：

$$B \cap D_d([x_i]_A,[x_j]_A)\neq\varnothing \tag{5.10}$$

为决策信息系统的决策辨识矩阵，从而得到系统的决策辨识矩阵。且 *B* 的任何真子集均不为决策协调集时，称 *B* 为决策约简集，即可以保留系统决策不变的约简属性集，约简后可以降低系统的冗余度。

从而确定影响恐怖袭击的主要属性，进而确定主要原因。

5.2.2 蔓延特性模型

恐怖袭击事件的蔓延特性指的是恐怖袭击事件如同传染病一样，会被效仿，造成类似传染病传播一样的效果，尤其是在对应国家的政府未对恐怖袭击事件进行任何的防范和打压措施的情形下，其蔓延性特征就越明显。恐怖袭击事件的蔓延特性与传染病极其相似，为此，本团队在借鉴传染病 SIRS 模型的基础上，提出一种恐怖袭击事件蔓延模型。旨在为预防和打压恐怖袭击事件提供一定理论基础。

（1）SIRS 模型

SIRS 模型是 SIR 模型的改进模型，SIRS 模型中的恢复状态的节点不会终身免疫，还会以一定的概率 λ 失去免疫力而转变为易感染节点，这类又会被感染节点所感染。因此 SIRS 模型用来描述免疫能力不足的传染病病毒的传播情况。这一情况与恐怖袭击事件蔓延特性极其吻合，恐怖势力在被打压后，并不是说以后就不会再爆发恐怖袭击事件了。对于恐怖袭击事件的打击、预防只有进行时，未有完成时，时一个长期坚持不懈的工作。

在 SIRS 模型中，所有节点也处于三种状态：易感染 Susceptible(S) 和感染状态 Infected(I)和免疫状态 Recovered(R)。处于易染状态的易染节点 S 与感染状态的感染节点 I 接触后会以一定的概率 β 转变成感染节点，感染节点通过药物或者自身免疫力以概率 μ 转变成恢复节点，同时也会有部分恢复节点由于免疫能力有限又会以一定概率 λ 转变成易染节点。SIRS 的传播示意图如图 5.3 所示：

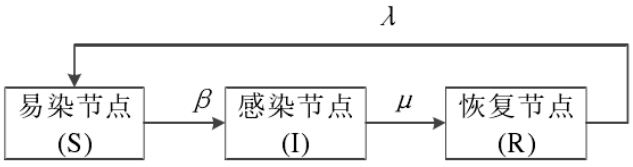


图 5.3 SIRS 传播模型图

这个过程可以用式(5.1)中的微分方程组进行描述，表示为：

$$\begin{cases} \dfrac{dS(t)}{dt} = \lambda S(t) - \beta I(t)S(t) \\ \dfrac{dI(t)}{d(t)} = \beta I(t)S(t) - \mu I(t) \\ \dfrac{dR(t)}{d(t)} = \mu I(t) - \lambda S(t) \end{cases} \tag{5.11}$$

其中，$S(t)$、$I(t)$、$R(t)$分别对应种群中 Susceptible(S)、Infected(I)、Recovered(R)三种个体在t时刻所占的比例，且有$S(t)+I(t)+R(t)=1$。式（5.11）中第一个式子表示易染节点的变化率，第二个式子表示感染节点的变化率，第三个式子表示恢复节点的变化率。

（2）恐怖袭击事件蔓延模型

本团队采用曲线拟合方法建立恐怖袭击事件蔓延模型。曲线拟合又称作函数逼近，是求近似函数的一类数值方法．它不要求近似函数在每个节点处与函数值相同，只要求其尽可能的反映给定数据点的基本趋势以及某种意义上的无限“逼近”．采用曲线拟合处理数据时，一般会考虑到误差的影响，于是我们往往基于残差的平方和最小的准则选取拟合曲线的方法，这便是经常所说的曲线拟合的最小二乘法．

➢ 最小二乘法

从整体上考虑近似函数$p(x)$同所给恐袭数据数据点(x_i,y_i) $(i=0,1,\ldots,m)$误差$r_i=p(x_i)-y_i$ $(i=0,1,\ldots,m)$的大小，常用的方法有以下三种：一是误差$r_i=p(x_i)-y_i$ $(i=0,1,\ldots,m)$绝对值的最大值$\max\limits_{0\le i\le m}|r_i|$，即误差向量$r=(r_0,r_1,...,r_m)^r$的∞—范数；二是误差绝对值的和$\sum\limits_{i=0}^{m}|r_i|$，即误差向量$r$的 1—范数；三是误差平方和$\sum\limits_{i=0}^{m}r_i^2$的算术平方根，即误差向量$r$的 2—范数；前两种方法简单、自然，但不便于微分运算 ，后一种方法相当于考虑 2—范数的平方，因此在曲线拟合中常采用误差平方和$\sum\limits_{i=0}^{m}r_i^2$来 度量误差$r_i$ $(i=0,1,\ldots,m)$的整体大小。

数据拟合的具体作法是：对给定恐袭数据(x_i,y_i) $(i=0,1,\ldots,m)$，在取定的函数类Φ中，求$p(x)\in\Phi$,使误差$r_i=p(x_i)-y_i$ $(i=0,1,\ldots,m)$的平方和最小，即

$$\sum_{i=0}^{m}r_i^2=\sum_{i=0}^{m}[p(x_i)-y_i]^2=min \tag{5.12}$$

从几何意义上讲，就是寻求与给定点(x_i,y_i) $(i=0,1,\ldots,m)$的距离平方和为最小的曲线☐$y=p(x)$（图 5.4）。函数$p(x)$称为拟合函数或最小二乘解，求拟合函数$p(x)$的方法称为曲线拟合的最小二乘法。

☐在曲线拟合中，函数类Φ可有不同的选取方法。

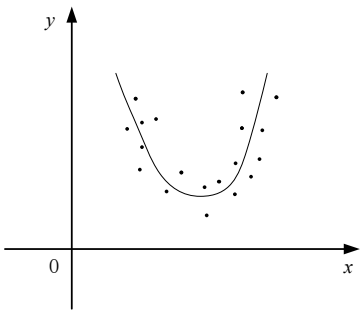


图 5.4 曲线 $y = p(x)$

➢ 多项式拟合□

假设给定恐袭数据的数据点 (x_i, y_i) (i=0,1,…， m)， Φ 为所有次数不超过 $n(n < m)$ 的多项式构成的函数类，现求 $p_n(x) = \sum_{k=0}^{n} a_k x^k \in \Phi$,使得

$$I = \sum_{i=0}^{m} [p_n(x_i) - y_i]^2 = \sum_{i=o}^{m} (\sum_{k=0}^{n} a_k x^k - y_i)^2 = min \tag{5.13}$$

当拟合函数为多项式时，称为多项式拟合，满足式(5.13)的 $p_n(x)$ 称为最小二乘拟合多项式。特别地，当 n=1 时，称为线性拟合或直线拟合。

显然 $I = \sum_{i=o}^{m} (\sum_{k=0}^{n} a_k x^k - y_i)^2$ 为 $a_0, a_1, ..., a_n$ 的多元函数，因此上述问题即为求 $I = I(a_0, a_1, ..., a_n)$ 的极值问题。由多元函数求极值的必要条件，得

$$\frac{\partial I}{\partial a_j} = 2\sum_{i=o}^{m} (\sum_{k=0}^{n} a_k x^k - y_i) x_i^j = 0, \quad j = 0,1,...,n \tag{5.14}$$

即

$$□ \sum_{i=o}^{m} (\sum_{k=0}^{n} x_i^{j+k}) a_k = \sum_{i=0}^{m} x_i^j y_i, \quad j = 0,1,...,n \tag{5.15}$$

式(5.15)是关于 $a_0, a_1, ..., a_n$ 的线性方程组，用矩阵表示为

$$\begin{bmatrix} m+1 & \sum_{i=0}^{m} x_i & \cdots & \sum_{i=0}^{m} x_i^n \\ \sum_{i=0}^{m} x_i & \sum_{i=0}^{m} x_i^2 & \cdots & \sum_{i=0}^{m} x_i^{n+1} \\ \vdots & \vdots & & \vdots \\ \sum_{i=0}^{m} x_i^n & \sum_{i=0}^{m} x_i^{n+1} & \cdots & \sum_{i=0}^{m} x_i^{2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{m} y_i \\ \sum_{i=0}^{m} x_i y_i \\ \vdots \\ \sum_{i=0}^{m} x_i^n y_i \end{bmatrix} \tag{5.16}$$

式(5.5)或式(5.6)称为正规方程组或法方程组。□

可以证明，方程组(5.16)的系数矩阵是一个对称正定矩阵，故存在唯一解。从式(5.16)中解出 a_k (k=0,1,…， n)，从而可得多项式□

$$p_n(x) = \sum_{k=0}^{n} a_k x^k \tag{5.17}$$

可以证明，式(5.17)中的 $p_n(x)$ 满足式(5.13)，即 $p_n(x)$ 为所求的拟合多项式。我们

把$\sum_{i=0}^{m}[p(x_i)-y_i]^2$称为最小二乘拟合多项式$p_n(x)$的平方误差，记作$\|r\|_2^2=\sum_{i=0}^{m}[p(x_i)-y_i]^2$，由式(5.14)可得□

$$\|r\|_2^2=\sum_{i=0}^{m}y_i^2-\sum_{k=0}^{n}a_k(\sum_{i=0}^{m}x_i^k y_i) \quad (5.18) \quad □$$

多项式拟合的方法可归纳为以下几步：

- 由已知数据画出函数粗略的图形——散点图，确定拟合多项式的次数 n；□
- 列表计算$\sum_{i=0}^{m}x_i^j \ (j=0,1,...,2n)$和$\sum_{i=0}^{m}x_i^j y_i \ (j=0,1,...,2n)$；□
- 写出正规方程组，求出$a_0,a_1,...,a_n$；
- 写出拟合多项式$p_n(x)=\sum_{k=0}^{n}a_k x^k$。

5.2.3 时空特性及级别分布模型

（1）恐怖袭击事件时空级别序列模型

恐怖袭击事件具有一定的时空特征，恐怖袭击事件随着国家政府的政权和社会状况的变化随时间呈规律性变化，同时在空间上也受到地区的影响，表现出在时间和空间上的连续型与渐变性。

考虑到恐怖袭击事件的空间特征，把恐怖袭击事件空间位置抽象为空间上的二维拓扑关系，建立恐怖袭击事件信息空间序列。空间序列根据时间序列的思想扩展而来。将某种随机变量按出现的时间顺序排序起来称为时间序列。时间序列是对某一个或一组变量 $x(t)$进行观察测量，将在一系列时刻$t_1,t_2,...,t_n$(t 为自变量且$t_1<t_2<...<t_n$)所得到的离散数字组成一个序列集合 $x(t_1)$，$x(t_2)$,…, $x(t_n)$，成为时间序列，可表示为：

$$x(t)=\{x(t_i),i=1,2,...,n\} \quad (5.19)$$

式(5.1)中，t_i表示恐怖袭击事件发生的日期，n 表示 $x(t)$ 序列中时间的个数。

根据时间序列的表示方式，空间序列可以表示为空间上连续或离散分布的对象间的一系列关系值的描述。对某一个或一组变量$x(s)$ 进行观察测量，s 表示地理位置，对一系列具有空间关系的对象$s_1,s_2,...,s_n$之间的联系的离散观测值组成的序列集合$x(s_{1,2})$，$x(s_{1,3}),...,x(s_{2,3}),x(s_{2,4}),\ldots,x(s_{n-1,n})$，把它称之为空间序列，表示如下：

$$x(s)=\{x(s_{i,j}),i=1,2,...,n;j=1,2,...,n;i\neq j\} \quad (5.20)$$

式中，n 为空间序列集合中包含的事件区域数；$s_{i,j}$ 则表示为这些位置间隐含的相互作用关系，$x(s_{i,j})$为位置 i 和位置 j 之间的联系值。

在问题一中本团队已经对恐怖袭击事件的级别进行分类，由于恐怖袭击事件的级别分布具有明显的时空特性，所以本团队在研究时空特性的同时也考虑了恐怖袭击事件的级别分布问题。r 表示恐怖袭击事件级别，则将事件级别加入到空间序列中去可以得到：

$$x(s,r)=\{x(s_{i,j},r_{i,j}),,i=1,2,...,n;j=1,2,...,n;i\neq j\} \quad (5.21)$$

根据上述方法建立起来的空间级别序列，结合恐怖袭击事件信息空间级别分布特点，根据事件发生不同地理位置之间的相互影响抽象出如下的空间级别序列模型，如图5.5所示。

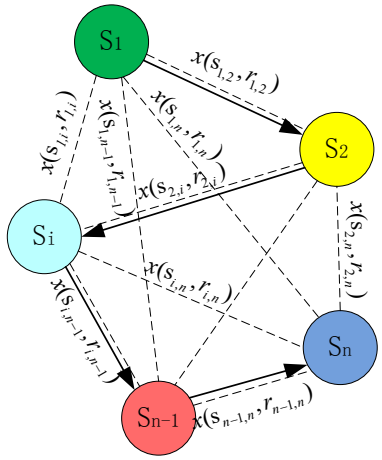


图5.5 恐怖袭击事件空间级别序列模型

时空级别序列建模分为在空间级别序列中插入时间序列和在时间序列中插入空间级别的方法。针对时空级别因素对恐怖袭击事件的重要作用，本团队采用把空间级别子序列作为时间序列的一个元素嵌入到时间序列中去，形成时空序列模型。根据式(5.19)和式(5.21)所示，在时间序列中嵌入空间级别序列，即是把时间序列中的每一个元素用空间级别序列来表示，空间级别序列中保留了各个空间对象和它们之间的联系值和事件级别，这些空间对象间的联系值随时间发生相应变化。把式(5.3)中的空间级别序列$x(s,r)$作为式(5.19)中时间序列$x(t)$的一个元素，则式(5.19)变为:

$$f(t)=\left\{t_i(\mathrm{s}_{j,k},r_{j,k}),,i=1,2,...,m;j,k=1,2,...,n;k\neq j\right\} \tag{5.22}$$

式中，$t_i(\mathrm{s}_{j,k},r_{j,k})$代表了在时间$i$时，空间级别对象$j$和$k$之间的联系值。

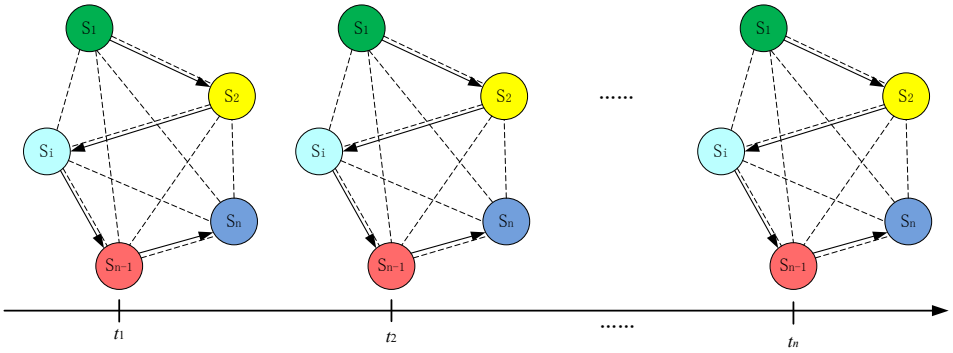


图5.6 恐怖袭击事件时空级别序列模型

本团队的研究对象是针对恐怖袭击事件的时空级别序列建模，本模型很好地表达了空间对象的空间级别关系和随时间变化的情况，反映了恐怖袭击事件在时间上、空间上和级别上动态变化和相互联系的特征，把恐怖袭击事件信息的时间、空间和级别属性信息融合到一起。

（2）RBF 神经网络模型

RBF神经网络是Moody和Darken于20世纪80年代末提出的一种具有单隐层的三层前馈网络。输入层由信号源节点组成;第二层是隐层，隐层的变换函数是RBF，它是对中心点对称且衰减的非负非线性函数;第三层为输出层，它对输入模式的作用做出响应。与其它前馈型神经网络相比，RBF神经网络具有良好的函数逼近性能，若RBF神经网络的隐

含层神经元个数足够多，则RBF神经网络可以一致连续逼近任何连续函数。利用RBF神经网络进行预测，首先要构建其网络模型，在建立RBF神将网络时，各层的节点数目、RBF、隐层中心、扩展常数和隐层到输出层的权值都是需要考虑的因素。把之前建立的时空级别序列模型结合RBF神经网络建立如下神经网络预测模型，如图5.7所示。

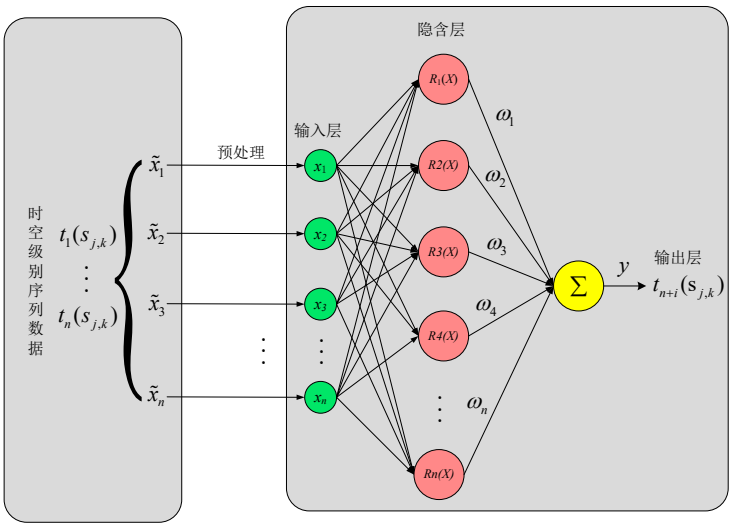


图5.7 基于时空级别序列的RBF神经网络模型

5.3 模型解算

5.3.1 基于粗糙集的恐怖袭击事件属性度量模型求解

由于在 3.3.1 和 3.3.2 两节中已经给出了属性权重的确定方法，并给出了详细的处理流程，此节不再赘述，只是得到结论。

即找到与疑似恐怖主义（doubtterr）相关的量，通过粗糙集处理可得主要为：

政治、经济、宗教或社会目标（crit1）

国家（country）

地区（region）

目标/受害者类型（targtype1）

犯罪集团的名称（gname）

总索取赎金（ransomamt）

从粗糙集处理的结果来看，上述属性权重较大，即为主要原因。

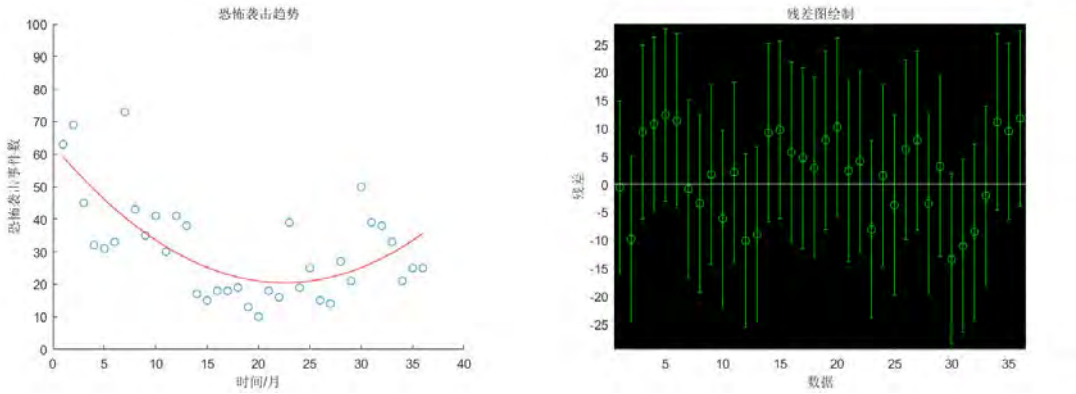
5.3.2 蔓延特性模型求解

本团队根据提出的蔓延特性模型对世界上危害性极大的恐怖组织之一——“博科圣地”(Boko Haram)进行分析。本团队首先统计了“博科圣地”近三年每个月的恐怖袭击事件数量，然后运用蔓延特性模型对“博科圣地”的恐怖袭击事件数量特征进行回归，得出“博科圣地”近三年来的恐怖袭击事件数量变化模型。模型求解结果能够为下一步分析研判各地区及世界的反恐态势提供依据

“博科圣地”近三年恐怖袭击事件数量模型及模型残差图如图 5.8 所示。图中横轴为从 2015 年 1 月开始到 2017 年 12 月三年的月数，纵轴为恐怖袭击事件数。求解的“博科圣地”恐怖袭击事件数量模型为：

$$y = 0.0866x^2 - 3.8726x + 62.41 \tag{5.23}$$

由求解的模型可以发现，“博科圣地”在 2015 年初每月的恐怖袭击事件数量较多。随着时间的推移，每月的恐怖袭击事件数量逐渐下降，至 2016 年底下降到最低水平，维持在每月 20 起事件左右。然而，2017 年初开始，虽然“博科圣地”的恐怖袭击事件数量与 2015 年初相比仍然处于较低水平，但开始有着缓慢增加的趋势。



(a)数量模型　　(b)残差

图 5.8 “博科圣地”恐怖袭击事件模型

结合因特网上关于“博科圣地”的相关信息，也程度上验证了我们求解的模型。2015 年开始，随着国际反恐势力的介入和尼日利亚政府军加大对“博科圣地”的打击力度，“博科圣地”确实遭受了巨大打击。2016 年底尼日利亚总统布哈里甚至宣布，极端组织“博科圣地”已经被摧毁。从本团队的“博科圣地”恐怖袭击事件蔓延模型求解结果可以发现，“博科圣地”在 2016 年确实受到了很大打击。本团队建立该蔓延模型是为了对恐怖势力未来的活动趋势进行预测，从蔓延模型求解结果发现，在刚刚过去的一年——2017 年里，“博科圣地”的活动有逐渐增加的趋势。通过“博科圣地”蔓延特性我们可以预测，极端组织“博科圣地”只是在 2016 年受到了很大的打击，但并没有被彻底摧毁且“博科圣地”正在逐渐恢复。2018 年“博科圣地”发动的恐怖袭击事件数量仍然会增加，需要加大针对“博科圣地”的反恐力度。

5.3.3 时空特性及级别分布模型求解

为了验证所构建的改进的 RBF 的准确性，将“博科圣地”2015 年 1 月至 2017 年 11 月共计 35 个月的事件汇总，得到各个事件发生点的坐标与等级，并按照时间顺序排序，用于训练 RBF 网络。网络迭代次数为 10 万次。学习效率为 0.3，目标梯度为 10^{-8} 进行训练。图 5.9 即为终止时，RBF 的误差曲线图。

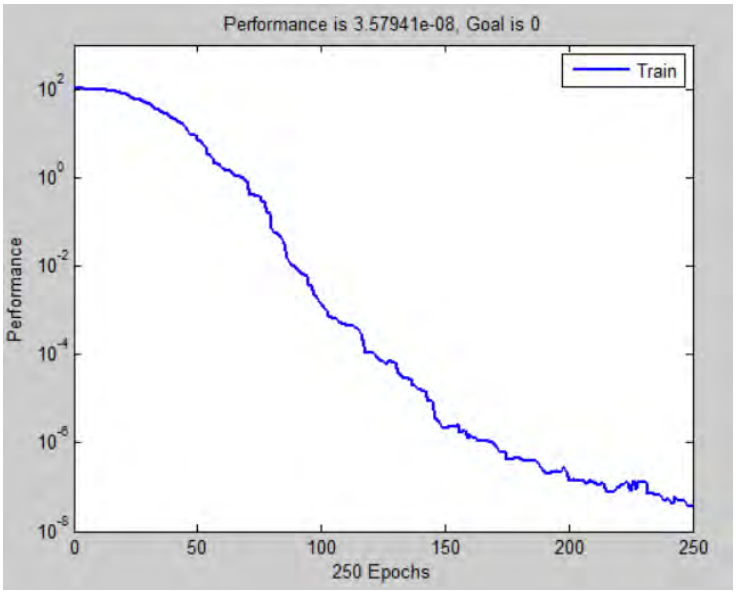


图 5.9 RBF 训练效果图

结合上一小结得到的事件次数，确定了次月可能发生 28 起事件，因此，将 11 月最后一个事件的坐标与评级带入训练好的网络中，得到第一个输出，再将此输出作为输入，重新带入网络中，得到第二组输出，重复第二代 28 次，得到 28 个坐标点与对应的事件分级。图 5.10 中红点表示预测地点，蓝点为实际恐怖主义事件发生的地点。

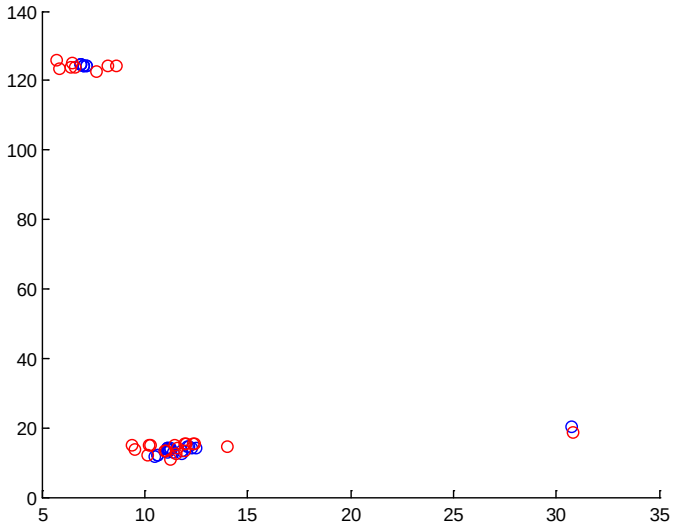


图 5.10 预测事件与真实事件的地理坐标

同样得到级别分布如图 5.11 所示。

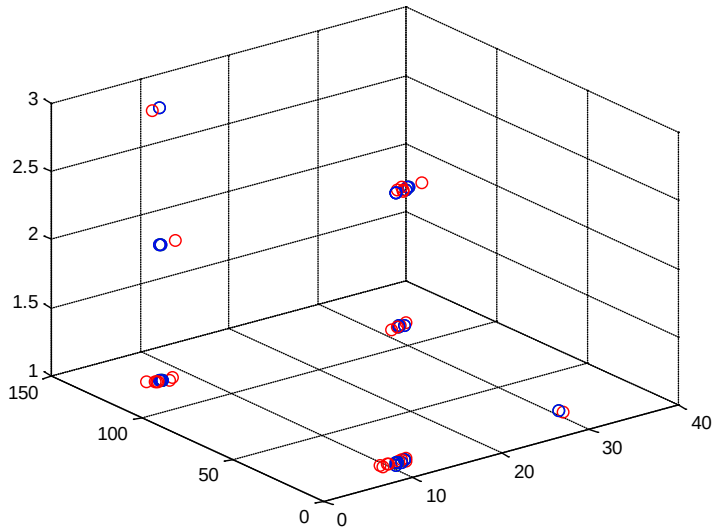


图 5.11 预测事件与真实事件的地理坐标-危害程度对比

通过上述仿真可以看出，本文方法具有一定的准确性与有效性，由于原始的恐怖组织具有一定的规律性，学习后能够很好的估计出可能发生恐怖主义袭击事件的地点与等级。

将某个区域多个具有代表性的组织通过上述方法进行预测，即可大致的描述该区域甚至全世界恐怖主义的分布态势。

5.4 结论及分析

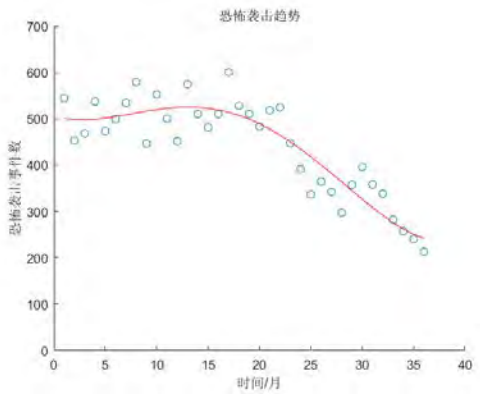
由附件 1 中数据我们可以发现，中东—北非、南亚和次撒哈拉非洲地区的恐怖袭击事件数占全世界恐怖袭击总数的 80%左右。且恐怖袭击事件跟其它地区相比单次事件的平均死亡人数较高，所以本团队确认中东—北非、南亚、次撒哈拉非洲三个地区为恐怖袭击重点地区。

- 中东—北非地区

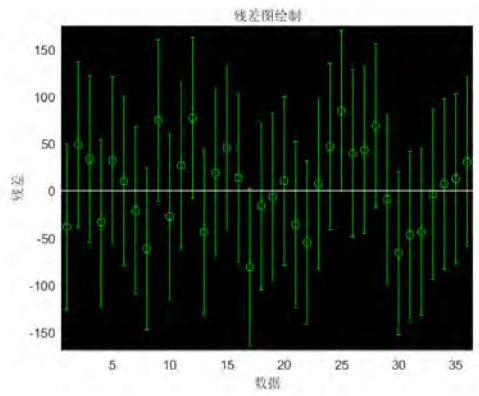
中东—北非恐怖袭击事件近三年数量模型及模型残差图如图 5.12 所示，求解的中东—北非恐怖袭击事件数量模型为：

$$y = -0.0014x^4 - 0.1004x^3 + 1.8229x^2 - 8.2978x + 507.91 \tag{5.24}$$

由求解的模型可知，2015-2016 年中东—北非地区的恐怖袭击数量变化不大，每月维持在 500 起左右。从 2017 年 1 月开始，恐怖袭击事件数量逐渐下降，到 2017 年底，降为 250 起左右。



(a)数量模型



(b)残差

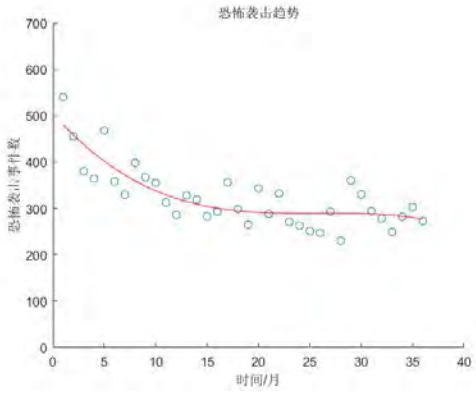
图 5.12　中东—北非恐怖袭击事件模型

➢ 南亚地区

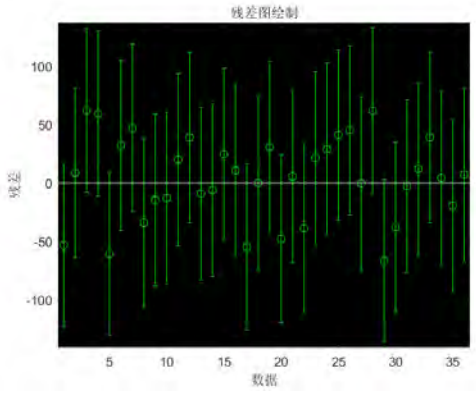
南亚地区恐怖袭击事件近三年数量模型及模型残差图如图 5.13 所示。求解的南亚恐怖袭击事件数量模型为：

$$y = -0.0144x^3 + 1.0935x^2 - 27.3378x + 515.55 \tag{5.25}$$

由求解的模型可知，从 2015 年 1 月开始，南亚地区恐怖袭击事件呈逐渐下降的趋势，到 2015 年底降到每月 300 起左右。2016-2017 年每月的恐怖袭击事件在 300 上下浮动。



(a)数量模型



(b)残差

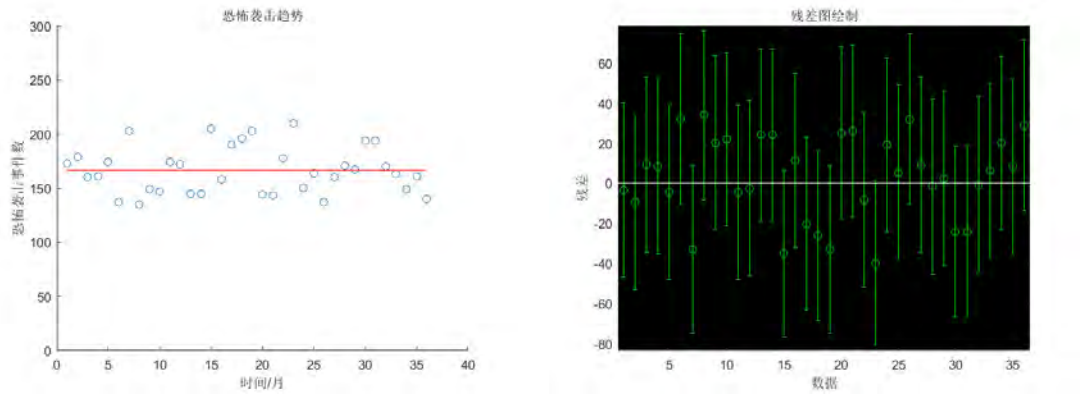
图 5.13　南亚恐怖袭击事件模型

➢ 次撒哈拉非洲地区

次撒哈拉非洲地区恐怖袭击事件近三年数量模型及模型残差图如图 5.14 所示。求解的南亚恐怖袭击事件数量模型为：

$$y = -0.0032x + 166.75 \tag{5.26}$$

由求解的模型可知，次撒哈拉非洲地区恐怖袭击事件在近三年始终维持在每月 160 起左右，说明在近几年中次撒哈拉非洲地区恐怖袭击事件维持在稳定水平。



(a)数量模型　　(b)残差

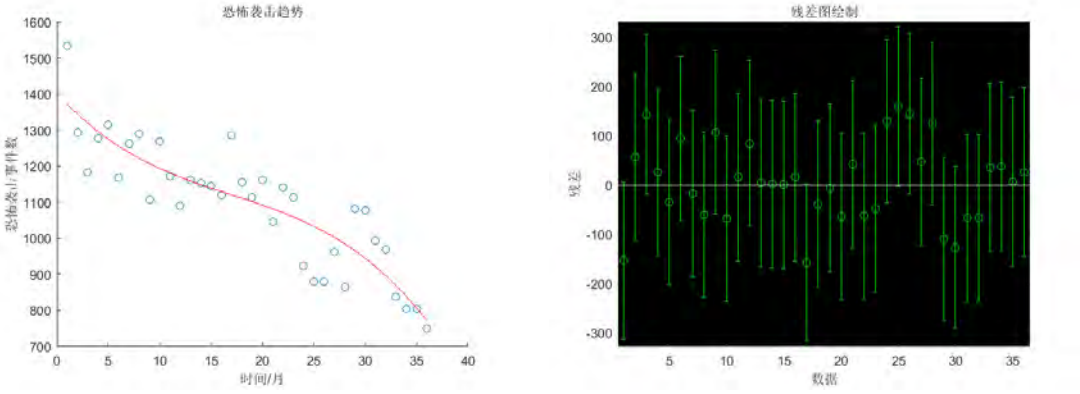
图 5.14　次撒哈拉非洲恐怖袭击事件模型

➢ 世界

从求解的恐怖袭击重点地区恐怖袭击事件数量模型可以预测世界的恐怖袭击事件数量变化趋势。由恐怖袭击重点地区求解的数量模型可知，世界恐怖袭击事件数量总体趋势呈现下降趋势。图 5.15 所示的世界恐怖袭击事件的数量模型显然验证了以上分析，求解的世界恐怖袭击事件数量模型为：

$$y = -0.0253x^3 + 1.2931x^2 - 31.2258x + 1401.9 \tag{5.27}$$

由求解的模型可以发现世界近几年的恐怖袭击事件逐渐减少，世界的反恐形势在总体上趋于好转。



(a)数量模型　　(b)残差

图 5.15 世界恐怖袭击事件模型

通过分析可知，从近三年的恐怖袭击事件发生的次数来看（详见图 5.16），虽然，全球范围内恐怖主义活动数量有所下降，但全球反恐形式依然不容乐观。我们重点分析了 5 个重点恐怖组织中的“博科圣地”和 ISIL 两个组织，从死亡人数来看，大多数死者来自发达国家，这跟“博科圣地”和 ISIL 两个组织一直试图扩大在其他国家和地区活动有一定的关系。那么为何 2015 年以后，这两个组织实施的恐怖袭击次数开始减少呢？本团队在通过查询相关背景资料和数值分析的基础上，发现外国军事力量对伊拉克和尼日利亚的武装干预，直接影响到了“博科圣地”和 ISIL 的主要活动范围，因此这两个国家死于恐怖主义活动的人数骤减了 20%。

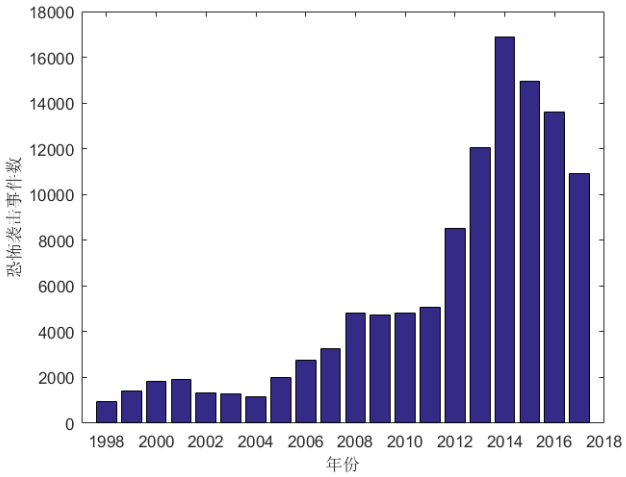


图 5.16 2015-2017 年全球恐怖袭击事件数量统计

从蔓延特性分析来看，这两个恐怖组织正向包括经合组织成员国在内的临近国家和地区，努力传播恐怖主义，从而提高其恐怖主义对世界其他地区的影响，因此，反恐形势依然不容乐观。另外，年轻人缺乏机会，对选举制度失去信任、犯罪率攀升和容易获得武器等社会经济因素上表现糟糕。这些社会经济因素更容易诱发恐怖袭击事件。为此，各国政府可以在这些方面着手，做好防恐措施。同时，在恐怖主义日趋全球化的状态下，国际社会必须加强反恐合作，在现实领域和互联网平台双管齐下，打击极端主义思想传播，并不断积极推动和促进联合国安理会发挥更大作用，在提高各国反恐能力基础上，真正形成国际反恐合力，遏制恐怖主义的发展和蔓延。

利用上述构建的时空序列与级别分布模型，并结合蔓延模型，对 2018 年 1 月的恐怖主义态势进行预测，结果如下图所示。

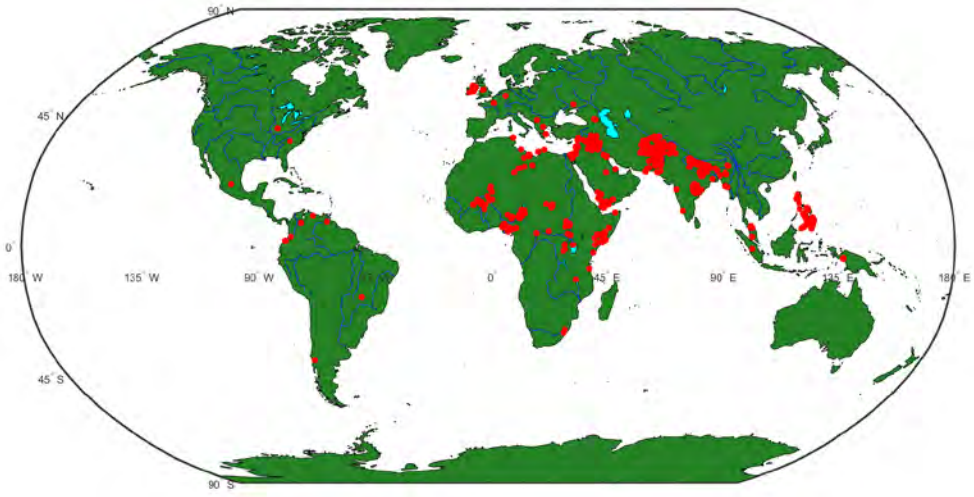


图 5.17 2018 年 1 月世界恐怖主义态势分布图

结论.：由 2018 年 1 月世界恐怖主义态势分布图分析可得，世界及各恐怖袭击重点地区的态势有向好趋势，但世界恐怖袭击事件仍然主要集中在中东—北非、南亚及次撒哈拉非洲等恐怖主义重灾区，且一些政权稳定、社会和谐的国家也不可掉以轻心，不排除发生恐怖主义袭击事件的可能性。

6 问题四的建模与求解

6.1 问题分析

本任务要求挖掘附件 1 中的信息，进一步发挥其作用。

在附件一中，有一类恐怖袭击事件，其并不发生在战火纷飞的区域，同时发生该事件的区域发生恐怖袭击事件的频率也不高。而此类区域一般为政权稳定，社会和谐的区域或国家，这与中国的社会现状极其相似，本组认为此类事件虽少，却具有很强的代表性。对我国反恐的预防具有积极意义。

因此，我们团队设置的任务就是从附件一的数据中找出此类的事件，从而对其数据进行分析，得到经验或者教训来促进我国的反恐工作。

这一类事件发生的在非恐怖袭击的高发地带，且发生该事件的区域，恐怖袭击事件发生与频率与危害程度均不高，在附件一中是一类离群的散点。为获取这类散点，可以利用数据挖掘中比较成熟的离群点检测来得到这些点，进而进行分析。

本任务采用离群点检测的方法，确定出这些离群点。这些点所处的区域的国家与社会状态，与我国现阶段更为贴近，研究此类点对应的恐怖袭击事件，分析其数据对于加强我国的反恐工作具有一定的指导作用。

6.2 模型建立

为考虑不同特征在不同聚簇中的作用，以及充分利用数据集中少量的先验知识，我们打算从聚类粒化的角度出发，借鉴文献[3]提出一种基于特征加权半监督聚类粒化的离群点挖掘方法（SSOD-FW）。已有的离群点挖掘方法同等看待每个特征，并没有考虑不同特征在每个聚簇中的不同重要性。根据特征的不同作用对不同的特征赋予权重，对无关特征赋予较低的权重，有利于解决高维稀疏数据的离群点挖掘问题。SSOD-FW 方法首先提出了一种基于特征加权半监督可能性 C-均值聚类的粒化策略，然后根据聚类诱导出的模糊粒结构定义数据点的离群程度。在聚类粒化过程中借助半监督指示矩阵将标记信息引入到新的聚类目标函数中，该目标函数为不同的特征分配自适应权重，综合考虑了聚类和离群点挖掘之间的相互影响，并假定遵循以下原则：

（1）最大化标记正常点对于其所属聚簇的隶属度；

（2）最小化标记正常点对于其非所属聚簇的隶属度；

（3）最小化标记离群点对每个聚簇的隶属度。

SSOD-FW 旨在通过特征加权的半监督聚类算法完成对数据集的信息粒化，从而诱导出关于数据集的一种更加合理、精确的模糊粒描述，在此基础上定义数据点的离群度，最终挖掘出数据集中的离群点。算法的流程如图 6.1 所示

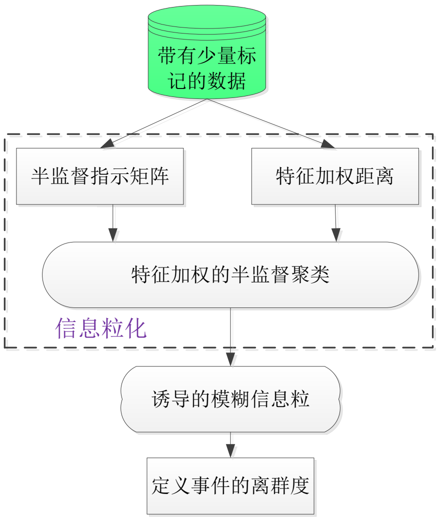


图 6.1 SSOD-FW 方法流程图

6.2.1 半监督指示矩阵的建立

设数据集 $X=\{x_1,x_2,...,x_t,x_{t+1},...,x_n\}$，其中前 t 个对象为未标记的数据，第 t+1 个到第 n 个对象为被标记的数据。由于聚簇内的点通常具有强相关性，离群点之间具有弱相关性，假设正常点属于 c 类中的某一个聚簇，而离群点不属于任何聚簇。定义了半监督指示矩阵（semi-supervised indicator matrix）$\mathbf{A}=(a_{il})_{n\times c}$。矩阵 $\mathbf{A}$ 中各元素的定义遵循如下规则：

（1）若样本 x_i 为标记的正常点并且 x_i 属于第 l 个聚簇，则 $a_{il}=-1$，并且对所有的 s =1,2,…,c，$s\neq l$，有 $a_{is}=1$。

（2）若样本 x_i 被标记为离群点，则对所有的 s=1,2,…,c，有 $a_{is}=1$。

（3）若样本 x_i 未被标记，则对所有的 s=1,2,…,c，有 $a_{is}=0$。

因此，得到 $\mathbf{A}$ 的具体表现形式如下：

$$\mathbf{A}=(a_{il})_{n\times c}=\left(\begin{array}{ccccccc} 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \hline -1 & 1 & \cdots & \cdots & \cdots & \cdots & 1 \\ \vdots & \vdots & \cdots & \cdots & \cdots & \ddots & \vdots \\ 1 & 1 & \cdots & -1 & 1 & \cdots & 1 \\ \vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ 1 & 1 & \cdots & \cdots & \cdots & 1 & -1 \\ \hline 1 & 1 & \cdots & \vdots & \cdots & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & \cdots & \cdots & \cdots & 1 \end{array}\right) \begin{array}{l} \text{未标记对象} \\ \text{标记的正常点} \\ \text{标记的离群点} \end{array}$$

6.2.2 特征加权距离

为充分考虑每个特征对不同聚簇的不同重要性和不同贡献，令 v_{jl} 为第 $j(1\leq j\leq N)$ 维特征在第 $l(1\leq l\leq c)$ 个聚簇中的权重，且满足 $\sum_{j=1}^{N}v_{jl}=1$。样本 x_i 到第 l 个聚类中心的特

征加权距离 d_{il} 考虑了每个特征在不同的聚簇中的不同重要性。d_{il} 定义为：

$$d_{il}=\sqrt{\sum_{j=1}^{N}(v_{jl})^{q}(x_{ij}-o_{ij})^{2}} \tag{6.1}$$

其中，$q>1$ 为预先给定的特征权值指数。

聚簇内部的样本点往往具有较强的相关性或相似性，然而离群点之间往往具有较弱的相关性。正常点对其所属聚簇的隶属度较大，离群点对于所有聚簇的隶属度均较小。通过特征加权距离和半监督指示矩阵 **A** 将特征权值和少量标记信息引入到基于聚类的离群点挖掘中，建立一个新的聚类粒化目标函数。该目标函数最大化标记正常点对其所属聚簇的隶属度，最小化标记正常点对其非所属聚簇的隶属度，以及最小化标记离群点对每个聚簇的隶属度。因此，SSOD-FW 算法的目标函数如下：

$$\begin{aligned}&\min J_{SSOD-FW}(U,V,O)=\sum_{i=1}^{n}\sum_{l=1}^{c}u_{il}^{m}d_{il}^{2}+\frac{\beta}{m^{2}\sqrt{c}}\sum_{i=1}^{n}\sum_{l=1}^{c}(u_{il}^{m}\log u_{il}^{m}-u_{il}^{m})+\sum_{i=1}^{n}\sum_{l=1}^{c}\alpha_{il}a_{il}u_{il}^{m}\\&s.t\quad u_{il}\in[0,1],\ \ v_{jl}\in[0,1],\ \ \sum_{j}^{N}v_{jl}=1\end{aligned} \tag{6.2}$$

其中，$\beta=\dfrac{\sum_{i=1}^{n}\sum_{j=1}^{N}(x_{ij}-\sum_{t=1}^{n}x_{tj}/n)^{2}}{n}$ 为数据集样本协方差；n、N 和 c 分别表示样本个数、特征个数和聚簇数目；$m(m>1)$ 为模糊因子；$U=(u_{il})_{n\times c}$ 为样本 x_i 对于第 l 个聚簇的隶属度；$O=(o_{lj})_{c\times N}$，o_{lj} 为第 l 个聚类中心的第 j 维特征；$V=(v_{jl})_{N\times c}$，v_{jl} 为第 j 维特征在第 l 个聚类中的权重；$a_{il}\in\{1,-1,0\}$ 表示半监督指示矩阵 **A** 中第 i 行 l 列的元素；参数 $\alpha_{il}(0<\alpha_{il}<1)$ 用于调节样本 x_i 隶属于第 l 个聚簇的先验信息在目标函数中发挥的作用，α_{il} 值越大，表示先验信息发挥的作用越大。

最小化目标函数 $J_{SSOD-FW}$ 第一项的目的在于最小化聚簇内部样本的特征加权距离，使得同一个聚簇内的数据散度最小，同一聚簇内的样本更加聚集。第二项要求隶属度 u_{il} 的值尽量大，避免平凡解的出现。目标函数通过第三项 $\sum_{i=1}^{n}\sum_{l=1}^{c}\alpha_{il}a_{il}u_{il}^{m}$ 引入先验信息，利用先验信息修正隶属度值，使得隶属度值更加准确。最小化目标函数第三项 $\alpha_{il}a_{il}u_{il}^{m}$，即最小化标注为离群点的样本对于每类的隶属度，最大化标注为正常点的样本对于所属类的隶属度。通过对 α_{il} 的合理取值，能够平衡的每个对象的标记信息在聚类过程中的权值，进而有望获得对数据集 X 的最优模糊划分。

6.3 模型解算

利用拉格朗日乘子法求解满足优化问题（6.2）的必要条件，分别推导 $\boldsymbol{V}$、$\boldsymbol{U}$ 和 $\boldsymbol{O}$ 的迭代更新公式。

（1）更新迭代特征权值 $\boldsymbol{V}$

首先固定 U 和 O，最小化 $J_{SSOD-FW}(V)$ 关于 V 的函数。在求解 V 的更新公式中，参数 $\alpha_{il}(1\le i\le n,1\le l\le c)$ 作为常数。设 $\lambda_l(l=1,2,...,c)$ 为拉格朗日因子，构建目标函数 $J_{SSOD-FW}$ 关于 $\boldsymbol{V}$ 的拉格朗日函数 $L(\boldsymbol{V})$ 如下：

$$L(V)=\sum_{i=1}^{n}\sum_{l=1}^{c}u_{il}^{m}d_{il}^{2}+\frac{\beta}{m^{2}\sqrt{c}}\sum_{i=1}^{n}\sum_{l=1}^{c}(u_{il}^{m}\log u_{il}^{m}-u_{il}^{m})+\sum_{i=1}^{n}\sum_{l=1}^{c}\alpha_{il}a_{il}u_{il}^{m}-\sum_{l=1}^{c}\lambda_{l}(\sum_{j=1}^{N}v_{jl}-1) \tag{6.3}$$

令 $L(V)$ 关于 v_{jl} 的偏导数为 0：

$$\frac{\partial L}{\partial v_{jl}} = q(v_{jl})^{(q-1)} \sum_{i=1}^{n} u_{il}^{m} (x_{ij} - o_{ij})^2 - \lambda_l = 0 \tag{6.4}$$

则

$$v_{jl} = \left[\frac{\lambda_l}{q \sum_{i=1}^{n} u_{il}^{m} (x_{ij} - o_{ij})^2} \right]^{1/(q-1)} \tag{6.5}$$

又由优化问题(6.2) 的约束条件得:

$$\sum_{j=1}^{N} v_{jl} = (\frac{\lambda_l}{q})^{1/(q-1)} \sum_{j=1}^{N} \left[\frac{1}{\sum_{i=1}^{n} u_{il}^{m} (x_{ij} - o_{ij})^2} \right]^{1/(q-1)} = 1 \tag{6.6}$$

即

$$(\frac{\lambda_l}{q})^{1/(q-1)} = \frac{1}{\sum_{j=1}^{N} \left[1 / \sum_{i=1}^{n} u_{il}^{m} (x_{ij} - o_{ij})^2 \right]^{1/(q-1)}} \tag{6.7}$$

将式（6.7）代入式（6.5）可知，特征权值 $v_{jl}(1 \le j \le N, 1 \le l \le c)$ 的迭代公式为：

$$v_{jl} = \frac{\left[1 / \sum_{i=1}^{n} u_{il}^{m} (x_{ij} - o_{ij})^2 \right]^{1/(q-1)}}{\sum_{p=1}^{N} \left[1 / \sum_{i=1}^{n} u_{il}^{m} (x_{ij} - o_{ip})^2 \right]^{1/(q-1)}} \tag{6.8}$$

特征权值 v_{jl} 的更新公式表明，在第 j 个特征空间中，假如数据在第 l 个聚簇中心周围分布比较紧致，则第 j 个特征具有较大的权值，在形成第 l 个聚簇中发挥较大的作用。对无关特征赋予较小的特征值，减弱了无关特征对挖掘过程的不利影响。

（2）迭代更新 O

为了求解聚簇中心矩阵 O，固定 U 和 V，参数 $\alpha_{il}(1 \le i \le n, 1 \le l \le c)$ 为常数。因为在优化问题（6.2）中，关于 O 没有限制条件。因此，求解目标函数 $J_{SSOD-FW}$ 关于 o_{lj} 的偏导数，并令其为 0：

$$\frac{\partial J_{SSOD-FW}}{\partial o_{lj}} = \sum_{i=1}^{n} (-2 u_{il}^{m} v_{jl} (x_{ij} - o_{ij})) = 0 \tag{6.9}$$

则聚类中心的更新规则为：

$$o_{lj} = \frac{\sum_{i=1}^{n} u_{il}^{m} x_{ij}}{\sum_{i=1}^{n} u_{il}^{m}} \tag{6.10}$$

（3）迭代更新 U

为求解模糊划分矩阵 U，固定 O 和 V，并且参数 $\alpha_{il}(1 \le i \le n, 1 \le l \le c)$ 为常数。求解 $J_{SSOD-FW}$ 关于 u_{il} 的偏导数，并令其为 0：

$$\frac{\partial J_{SSOD-FW}}{\partial u_{ij}} = m u_{il}^{m-1} d_{il}^2 + \frac{\beta}{m^2\sqrt{c}}(m u_{il}^{m-1}\log u_{il}^m) + \alpha_{il} a_{il} u_{il}^{m-1} = 0 \tag{6.11}$$

则 U 的更新公式为：

$$u_{il} = \exp\left[\frac{-m\sqrt{c}}{\beta}(d_{il}^2 + \alpha_{il} a_{il})\right] \tag{6.12}$$

式（6.12）表明，对于任意$1 \le i \le n$和$1 \le l \le c$，若d_{il}值较大，则相应的u_{il}值较小。同时，隶属度u_{il}的取值依赖于α_{il}的值。α_{il}用于度量标记信息在聚类过程中的重要程度，故α_{il}值的选择是影响 SSOD-FW 方法粒化效果的关键问题。如果α_{il}值太小，目标函数$J_{SSOD-FW}$第三项将被忽略，先验信息在聚类粒化过程中不能发挥作用。如果α_{il}值太大，目标函数$J_{SSOD-FW}$的第一项和第二项将被忽略，先验信息中若存在错误的标记信息，其对聚类过程的负面影响将会被放大，严重影响聚类效果。因此，α_{il}的合理取值应该与$J_{SSOD-FW}$的第一项同阶。这里，采用自适应的方法确定参数α_{il}，后续的实例验证部分令：$\alpha_{il} = K d_{il}^2$，其中，参数$K(0 < K < 1)$调节标记信息对隶属度的影响。K值越大，已知标注信息对离群点挖掘的影响越大。因为d_{il}是动态更新的，参数α_{il}在每步迭代中也需动态更新。

6.4 实例验证及结果分析

本团队利用提出的 SSOD-FW 方法进行离群点挖掘。首先统计 2017 年 1 月发生的共 868 起恐怖袭击事件的经纬度坐标，然后利用提出的模型进行离群点挖掘，挖掘结果如图 6.2 所示。

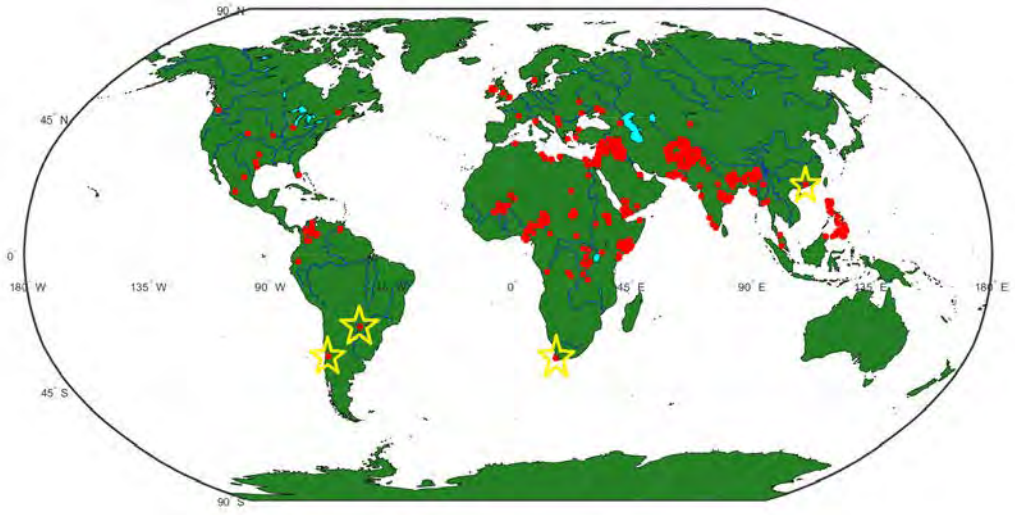
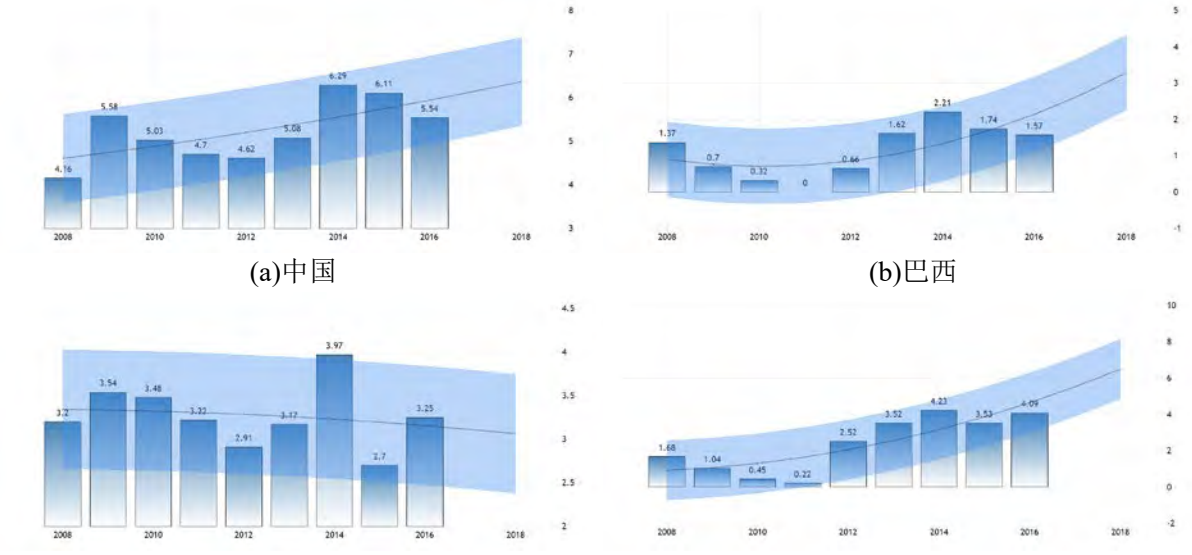


图 6.2 恐怖袭击事件离群点检测结果

图中黄色五角星符号标注的四个点为本团队模型挖掘的离群点。由图中四个点的分布来看，四个点分别在中国、南非、巴西和智利。这四个国家与中东—北非、南亚、次撒哈拉非洲地区国家不同，中国、南非和巴西均为金砖国家，而智利已经跳出中等收入陷阱，迈入了高收入国家行列。四个国家的政权相对稳固，且经济形势向好，社会较为稳定。四国恐怖主义指数预测如图 6.3 所示，通过恐怖主义指数比较发现中国、南非和

巴西三国的恐怖主义指数非常相似。这些国家发生恐怖袭击事件将会有明显区别于恐怖袭击重灾区国家，研究这些国家的恐怖袭击事件的共性特征以及反恐措施将具有一定的借鉴意义。



(a)中国　　(b)巴西

(c)智利　　(b)南非

图 6.3 四国近年恐怖主义指数及预测趋势

从这类恐怖袭击事件离群点所在国家的反恐形势不难得出以下共同点：

◆ 政府防恐应对体制单一

这类国家一般存在防恐主体不明确及事前预警、事中应对、事后救治缺乏系统化的问题。因为恐怖袭击事件离群点国家一般政权比较稳定，国家处于经济上升期甚至高收入国家行列，反恐经验不足，容易出现各部门职责不明确与反恐体制不完善问题。

◆ 社会福利制度不完善

随着国家的发展，此类国家的贫富差距逐渐拉大，而政府致力于缩小贫富差距的社会福利制度不够完善。弱势群体与社会财富的累积的落差逐渐增大，容易使其产生报复社会的心理，他国的恐怖势力也会随之趁机而入。

◆ 政府防恐与国际合作有限

目前的国际合作区域都是以联合国为主导的国际合作所构建的国际反恐机制，虽然对打击国际恐怖犯罪有一定的制约，但是并没有从根本上解决国际恐怖活动的发展和蔓延。由于各国之间文化政治经济都存在着差异，因此对恐怖主义的认识存在着不同的看法，这就导致在与国际合作时具有很大难度。

针对以上共性问题，我国可以借鉴相似恐怖袭击事件离群点所在国家反恐经验。首先，建立健全政府反恐应对机制系统，建立以政府为主导的应对机制，完善反恐应急机制；其次，完善社会福利制度，帮助弱势群体改善自己的生活水平；最后，建立多边双边国际合作。在双边多边合作领域，我国既要加强和一些大国之间的反恐合作，也要加强与相似恐怖袭击事件离群点所在国家的反恐交流。

参考文献

[1] 王锂达. 恐怖组织行为挖掘与预测[D]. 北京: 北京邮电大学, 2017.
[2] 莫豪文. 数据挖掘方法在反恐预警中的应用[D]. 北京: 北京工业大学, 2017.
[3] 杨金鸿. 基于粒计算的离群挖掘方法研究[D]. 哈尔滨, 哈尔滨工程大学, 2017.
[4] 孟祥茂. 基于复杂网络的 SIR 模型扩展与应用研究[D]. 南昌: 江西理工大学, 2015.
[5] 薛安荣. 空间离群点挖掘技术的研究[D]. 苏州, 江苏大学, 2008.
[6] 王振洲. 离群点检测方法研究及其在机器学习中的应用[D]. 武汉, 中国地质大学, 2018.
[7] 邱敦国, 兰时勇, 杨红雨. 基于时空特性的短时交通流预测模型[J]. 华南理工大学学报（自然科学版）, 2014, 42(7):49-54.
[8] 柴瑞瑞, 刘德海, 陈静锋. 恐怖袭击事件的时空差异特征分析及内生性 VAR 模型[J]. 中国管理科学, 2016, 24(专辑):281-288.
[9] 郭旺佳. 当代恐怖袭击的特点及中国的应对措施[D]. 上海: 上海国际问题研究院, 2017.

.

附　录

附录 1：

聚类：

```
[X,textdata] = uiopen('C:\Users\2016\Desktop\附件 1\附件 1.xlsx',1);
Taverage = clusterdata(X,'linkage','average','maxclust',3);
obslabel(Taverage == 1)
obslabel(Taverage == 2)
obslabel(Taverage == 3)
y = pdist(X);
Z = linkage(y,'average')
obslabel = textdata(2:end,1);
H = dendrogram(Z,0,'orientation','right','labels',obslabel);
set(H,'LineWidth',2,'Color','k');
xlabel('标准化距离（类平均法）')
inconsistent0 = inconsistent(Z,40)
```

附录 2：

拟合：

```
clear
clc
close all
uiopen('C:\Users\2016\Desktop\附件 1\附件 1.xlsx',1);
scatter(1:36,a);
hold on
y1 = polyfit(1:36,a',2);
x = 1:36;
y = y1(1)*x.^2+y1(2)*x.^1+y1(3);
plot(x,y,'r');
axis([0 40 0 100]);
title('恐怖袭击趋势');
xlabel('时间/月');
ylabel('恐怖袭击事件数');
figure
[b,bint,r,rint,stats]=regress(y',a);
rcoplot(r,rint);
title('残差图绘制');
xlabel('数据');
ylabel('残差');
```

附录 3：

RBF:

```
%% 清空环境变量
```

```
clc
clear

ld=400;

x=rand(2,ld);

x=(x-0.5)*1.5*2;

x1=x(1,:);
x2=x(2,:);

F=20+x1.^2-10*cos(2*pi*x1)+x2.^2-10*cos(2*pi*x2);
net=newrb(x,F);
interval=0.1;
[i, j]=meshgrid(-1.5:interval:1.5);
row=size(i);
tx1=i(:);
tx1=tx1';
tx2=j(:);
tx2=tx2';
tx=[tx1;tx2];

ty=sim(net,tx);

%% 使用图像，画出 3 维图

% 真正的函数图像
interval=0.1;
[x1, x2]=meshgrid(-1.5:interval:1.5);
F = 20+x1.^2-10*cos(2*pi*x1)+x2.^2-10*cos(2*pi*x2);
subplot(1,3,1)
mesh(x1,x2,F);
zlim([0,60])
title('真正的函数图像')

% 网络得出的函数图像
v=reshape(ty,row);
subplot(1,3,2)
mesh(i,j,v);
zlim([0,60])
title('RBF 神经网络结果')
```

```
% 误差图像
subplot(1,3,3)
mesh(x1,x2,F-v);
zlim([0,60])
title('误差图像')

set(gcf,'position',[300 ,250,900,400])
```

附录 4：

SVM 1：：

```
function chapter_sh
tic;
close all;
clear;
clc;
format compact;
uiopen('C:\Users\2016\Desktop\附件 1\附件 1.xlsx',1)

[m,n] = size(sh);
ts = sh(2:m,1);
tsx = sh(1:m-1,:);
figure;
plot(ts,'LineWidth',2);
grid on;
ts = ts';
tsx = tsx';
[TS,TSps] = mapminmax(ts,1,2);
figure;
plot(TS,'LineWidth',2);
grid on;

TS = TS';


[TSX,TSXps] = mapminmax(tsx,1,2);

TSX = TSX';



[bestmse,bestc,bestg] = SVMcgForRegress(TS,TSX,-8,8,-8,8);
[bestmse,bestc,bestg] = SVMcgForRegress(TS,TSX,-4,4,-4,4,3,0.5,0.5,0.05);
```

```
cmd = ['-c ', num2str(bestc), ' -g ', num2str(bestg) , ' -s 3 -p 0.01'];
model = svmtrain(TS,TSX,cmd);

[predict,mse] = svmpredict(TS,TSX,model);
predict = mapminmax('reverse',predict',TSps);
predict = predict';


figure;
hold on;
plot(ts,'-o');
plot(predict,'r-^');
legend('原始数据','回归预测数据');
hold off;
grid on;

figure;
error = predict - ts';
plot(error,'rd');
grid on;

figure;
error = (predict - ts')./ts';
plot(error,'rd');

grid on;
snapnow;
toc;
```

附录 5：

SVM 2：：

```
function chapter_FIGsh
tic;
close all;
clear;
clc;
format compact;

uiopen('C:\Users\2016\Desktop\附件 1\附件 1.xlsx',1)


ts = sh_open;
```

```
time = length(ts);


figure;
plot(ts,'LineWidth',2);
snapnow;
win_num = floor(time/5);
tsx = 1:win_num;
tsx = tsx';
[Low,R,Up]=FIG_D(ts','triangle',win_num);
figure;
hold on;
plot(Low,'b+');
plot(R,'r*');
plot(Up,'gx');
hold off;
grid on;
snapnow;
[low,low_ps] = mapminmax(Low);
low_ps.ymin = 100;
low_ps.ymax = 500;
[low,low_ps] = mapminmax(Low,low_ps);
figure;
plot(low,'b+');
grid on;
low = low';
snapnow;
[bestmse,bestc,bestg] = SVMcgForRegress(low,tsx,-10,10,-10,10,3,1,1,0.1,1);
[bestmse,bestc,bestg] = SVMcgForRegress(low,tsx,-4,8,-10,10,3,0.5,0.5,0.05,1);
cmd = ['-c ', num2str(bestc), ' -g ', num2str(bestg) , ' -s 3 -p 0.1'];
low_model = svmtrain(low, tsx, cmd);
[low_predict,low_mse] = svmpredict(low,tsx,low_model);
low_predict = mapminmax('reverse',low_predict,low_ps);
predict_low = svmpredict(1,win_num+1,low_model);
predict_low = mapminmax('reverse',predict_low,low_ps);
predict_low
figure;
hold on;
plot(Low,'b+');
plot(low_predict,'r*');
grid on;
figure;
error = low_predict - Low';
plot(error,'ro');
```

```
grid on;
[r,r_ps] = mapminmax(R);
r_ps.ymin = 100;
r_ps.ymax = 500;

[r,r_ps] = mapminmax(R,r_ps);

figure;
plot(r,'r*');
title('r 归一化后的图像','FontSize',12);
grid on;

[bestmse,bestc,bestg] = SVMcgForRegress(r,tsx,-4,8,-10,10,3,0.5,0.5,0.05);
cmd = ['-c ', num2str(bestc), ' -g ', num2str(bestg) , ' -s 3 -p 0.1'];
r_model = svmtrain(r, tsx, cmd);
[r_predict,r_mse] = svmpredict(r,tsx,low_model);
r_predict = mapminmax('reverse',r_predict,r_ps);
predict_r = svmpredict(1,win_num+1,r_model);
predict_r = mapminmax('reverse',predict_r,r_ps);
predict_r
figure;
hold on;
plot(R,'b+');
plot(r_predict,'r*');
legend('original r','predict r',2);
title('original vs predict','FontSize',12);
grid on;
figure;
error = r_predict - R';
plot(error,'ro');
title('误差(predicted data-original data)','FontSize',12);
grid on;
[up,up_ps] = mapminmax(Up);
up_ps.ymin = 100;
up_ps.ymax = 500;
[up,up_ps] = mapminmax(Up,up_ps);
figure;
plot(up,'gx');
title('Up 归一化后的图像','FontSize',12);
grid on;
up = up';
snapnow;
[bestmse,bestc,bestg] = SVMcgForRegress(up,tsx,-10,10,-10,10,3,1,1,0.5);
[bestmse,bestc,bestg] = SVMcgForRegress(up,tsx,-4,8,-10,10,3,0.5,0.5,0.2);
```

```
cmd = ['-c ', num2str(bestc), ' -g ', num2str(bestg) , ' -s 3 -p 0.1'];
up_model = svmtrain(up, tsx, cmd);
[up_predict,up_mse] = svmpredict(up,tsx,up_model);
up_predict = mapminmax('reverse',up_predict,up_ps);
predict_up = svmpredict(1,win_num+1,up_model);
predict_up = mapminmax('reverse',predict_up,up_ps);
predict_up
figure;hold on;
plot(Up,'b+');
plot(up_predict,'r*');
legend('original up','predict up',2);
title('original vs predict','FontSize',12);
grid on;
figure;
error = up_predict - Up';
plot(error,'ro');
title('误差(predicted data-original data)','FontSize',12);
grid on;
toc;
```

附录 6：

SVM 子程序：：

```
function [mse,bestc,bestg] = SVMcgForRegress(train_label,train,cmin,cmax,gmin,gmax,v,cstep,gstep,msestep)
%SVMcg cross validation by faruto

%
% by faruto
%Email:patrick.lee@foxmail.com QQ:516667408 http://blog.sina.com.cn/faruto BNU
%last modified 2010.01.17
%Super Moderator @ www.ilovematlab.cn

% 若转载请注明：
% faruto and liyang , LIBSVM-farutoUltimateVersion
% a toolbox with implements for support vector machines based on libsvm, 2009.
% Software available at http://www.ilovematlab.cn
%
% Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for
% support vector machines, 2001. Software available at
% http://www.csie.ntu.edu.tw/~cjlin/libsvm

% about the parameters of SVMcg
if nargin < 10
    msestep = 0.06;
```

```
    end
    if nargin < 8
        cstep = 0.8;
        gstep = 0.8;
    end
    if nargin < 7
        v = 5;
    end
    if nargin < 5
        gmax = 8;
        gmin = -8;
    end
    if nargin < 3
        cmax = 8;
        cmin = -8;
    end
    % X:c Y:g cg:acc
    [X,Y] = meshgrid(cmin:cstep:cmax,gmin:gstep:gmax);
    [m,n] = size(X);
    cg = zeros(m,n);

    eps = 10^(-4);

    bestc = 0;
    bestg = 0;
    mse = Inf;
    basenum = 2;
    for i = 1:m
        for j = 1:n
            cmd = ['-v ',num2str(v),' -c ',num2str( basenum^X(i,j) ),' -g ',num2str( basenum^Y(i,j) ),' -s 3 -p 0.1'];
            cg(i,j) = svmtrain(train_label, train, cmd);

            if cg(i,j) < mse
                mse = cg(i,j);
                bestc = basenum^X(i,j);
                bestg = basenum^Y(i,j);
            end

            if abs( cg(i,j)-mse )<=eps && bestc > basenum^X(i,j)
                mse = cg(i,j);
                bestc = basenum^X(i,j);
                bestg = basenum^Y(i,j);
            end
```

```
        end
    end
    % to draw the acc with different c & g
    [cg,ps] = mapminmax(cg,0,1);
    figure;
    [C,h] = contour(X,Y,cg,0:msestep:0.5);
    clabel(C,h,'FontSize',10,'Color','r');
    xlabel('log2c','FontSize',12);
    ylabel('log2g','FontSize',12);
    firstline = 'SVR 参数选择结果图(等高线图)[GridSearchMethod]';
    secondline = ['Best c=',num2str(bestc),' g=',num2str(bestg), ...
        ' CVmse=',num2str(mse)];
    title({firstline;secondline},'Fontsize',12);
    grid on;

    figure;
    meshc(X,Y,cg);
    % mesh(X,Y,cg);
    % surf(X,Y,cg);
    axis([cmin,cmax,gmin,gmax,0,1]);
    xlabel('log2c','FontSize',12);
    ylabel('log2g','FontSize',12);
    zlabel('MSE','FontSize',12);
    firstline = 'SVR 参数选择结果图(3D 视图)[GridSearchMethod]';
    secondline = ['Best c=',num2str(bestc),' g=',num2str(bestg), ...
        ' CVmse=',num2str(mse)];
    title({firstline;secondline},'Fontsize',12);
```

附录 7：

```
    TOPSIS：
    clear all
    clc
    uiopen('C:\Users\2016\Desktop\附件 1\（排序）附件 1.xlsx',1)

    wa=w.*a;
    aa=[1.0906 1.044 0.564 0.752];
    bb=[0 0 0.188 0.188];

    for i=1:10

s1(i)=(aa(1)-wa(i,1)).^0.5+(aa(2)-wa(i,2)).^0.5+(aa(3)-wa(i,3)).^0.5+(aa(4)-wa(i,4)).^0.5;

s2(i)=(wa(i,1)-bb(1)).^0.5+(wa(i,2)-bb(2)).^0.5+(wa(i,3)-bb(3)).^0.5+(wa(i,4)-bb(4)).^0.5;
```

```
end
s1=s1';
s2=s2';
   s1=abs(s1);

for i=1:10
       t(i)=s2(i)./(s1(i)+s2(i));
end
t=t';
```