情报杂志
Journal of Intelligence
ISSN 1002-1965,CN 61-1167/G3

《情报杂志》网络首发论文

题目：基于恐怖袭击特征分析的恐怖组织预测方法研究
作者：罗维平，周博
网络首发日期：2020-10-21
引用格式：罗维平，周博．基于恐怖袭击特征分析的恐怖组织预测方法研究[J/OL]．情报杂志. https://kns.cnki.net/kcms/detail/61.1167.G3.20201020.1701.004.html

网络首发时间：2020-10-21 09:53:08
网络首发地址：https://kns.cnki.net/kcms/detail/61.1167.G3.20201020.1701.004.html

基于恐怖袭击特征分析的恐怖组织预测方法研究

罗维平[1,2]　周　博[1]

（1. 武汉纺织大学 机械工程与自动化学院　武汉　430200；
2. 湖北省数字化纺织装备重点实验室　武汉　430200）

摘　要：[**目的/意义**]根据不同恐怖组织所具有的不同的特征构建恐怖活动与恐怖组织关联性预测模型，对凶手未知的恐怖活动嫌疑人进行预测，得到准确率及精度较高的预测结果。[**方法/过程**]在对 GTD 提供的恐怖袭击数据进行清洗筛选的基础上，使用经过特征工程选择且通过学习曲线验证能使模型达到最优效果的 29 个特征变量作为输入变量输入机器学习模型进行训练，并针对各模型特点进行模型结果融合得到最终关联性模型。[**结果/结论**]单个模型预测效果最佳的为随机森林，精度为 83.24%；对恐怖组织样本进行频率细分后，结合随机森林及 KNN 模型自身特点分别在高频、中频、低频段得到的模型精度为 88.53%、87.25%、86.14%，模型性能整体得到提升，特别在低频恐怖组织预测上也能得到较好表现。

关键词：恐怖袭击；预测模型；特征工程；机器学习；随机森林；KNN
中图分类号：TP181

Research on Terrorist Organization Forecast Method Based on Terrorist Attack Characteristics Analysis

Luo Weiping[1,2]　Zhou Bo[1]

(1. School of Mechanical Engineering and Automation, Wuhan Textile University, Wuhan　430200;
2. Hubei Digital Textile Key Laboratory of Digital Equipment, Wuhan　430200)

Abstract: [**Purpose/Significance**] According to the different characteristics of different terrorist organizations, a prediction model of the relevance between terrorist activities and terrorist organizations is constructed to predict terrorist suspects whose identity as murderers are unknown, and the prediction results with high accuracy and precision are obtained. [**Method/Process**] Based on the cleaning and screening of the terrorist attack data provided by GTD, 29 feature variables selected by feature engineering and verified by the learning curve to make the model achieve the best effect are used as input variables to input the machine learning model for training, and the model characteristics are fused with the model results to obtain the final correlation model. [**Result/Conclusion**] The best prediction effect of a single model is random forest, with an accuracy of 83.24%; after subdividing the frequency of terrorist organization samples, combined with the characteristics of random forest and KNN model, the accuracy of the model obtained at high frequency, intermediate frequency, and low frequency band is 88.53%, 87.25%, 86.14% respectively, the overall performance of the model has been improved, especially in the prediction of low-frequency terrorist organizations.

Key words: terrorist attack; prediction model; feature engineering; machine learning; random forest; KNN

0　引　言

近年来，世界范围内恐怖主义袭击事件发生频繁，加之全球信息化及媒体媒介的迅速发展，恐怖主义袭击事件所造成的社会影响日趋恶劣，根据全球恐怖主义数据库（Global Terrorism Database，GTD）所记载的

作者简介：罗维平（ORCID：0000-0003-3993-8090），女，1967 年生，硕士，教授，研究方向：大数据处理、检测技术与自动控制；周　博（ORCID：0000-0002-1441-2546），男，1997 年生，硕士研究生，研究方向：大数据处理、机器学习。

1998-2018 年世界范围内所有恐怖袭击数据显示，截止 2018 年 12 月 31 日，全球发生的恐怖袭击事件达 191464 起。恐怖袭击事件的发生对国家安全和地区稳定造成了很大的影响。

在所有恐怖袭击事件中，有绝大部分事件对于造成恐怖袭击的恐怖组织信息是未知的，使各国打击围剿行动受阻，因此通过对恐怖活动特征与恐怖集团关联性的研究，构建恐怖组织预测模型对凶手未知的恐怖袭击事件嫌疑人进行预测具有重要的现实意义，恐怖组织预测模型的建立可为打击恐怖组织犯罪提供更加准确的情报。本文通过特征相关性分析加以机器学习中几种分类算法对比融合，训练得到恐怖集团特征与恐怖组织关联性预测模型，为恐怖活动嫌疑人提供精度较为理想的预测结果。

1 文献综述

由于不同恐怖组织所造成的恐怖袭击事件具有不同的袭击特点，因此这些特征具有一定的指向性，所以对恐怖组织特征的研究是构建恐怖组织预测模型的前提。目前国内外学者针对恐怖集团的研究大多趋于定性研究，即针对单个或同一类型恐怖集团或是某一地区恐怖组织的定性分析。例如舒洪水等通过对 GTD 中 7133 次袭击样本进行统计分析得出“伊斯兰国”恐怖袭击的特点为：区域性聚集，全球化扩散；宗教性伪装，极端化本质；信息化串联，网络化扩张[1]。纳兰星舟等对“独狼”式恐怖袭击特征进行研究得出其袭击具有独立性、突发性和高效性、跨国性和报复性，主要针对“软目标”，防控难度大等特征[2-5]。白海娟等依据新疆历史、地理位置因素和宗教信仰因素对新疆地区恐怖活动特征做出相应分析，陈帅等依据空间统计和核密度分析方法分析了中南半岛恐怖袭击的时空演变特征，探讨了恐怖袭击目标和方式的转变特征[6-7]。

在恐怖活动特征研究应用方面，目前国内外众多学者根据恐怖活动特征对恐怖袭击事件嫌疑人预测进行了相关研究。A. Xue 等提出一种基于上下文子空间的预测算法 PBCS 对恐怖行为进行预测，但其缺陷在于如果无法检测到所有的属性，就会导致算法的波动，降低预测精度[8]。Nurudeen. M 等根据模拟广域监视网络中提取的犯罪指示事件提出一种混合神经模糊模型，来预测广泛地区的犯罪行为[9]；Li. Z 等提出了一种结合社交网络分析，小波变换及模式识别方法的综合框架，对恐怖组织行为进行预测[10]；战兵等利用隐马尔可夫模型与贝叶斯网络方法构建了恐怖事件预测模型[11]；Lin. Y. L 等基于破窗理论和空间分析的数据驱动方法，利用机器挖掘算法对犯罪数据进行分析，对犯罪热点进行了预测；李慧等提出了一种基于机器学习方法的预测模型，其对单个嫌疑人预测准确率较高，但对于多起恐怖事件的嫌疑人预测效果较低[12-13]。通过对上述模型的研究得出由于建模所用数据量较少，恐怖集团特征选择方法单一等原因造成了各模型未达到较高的精确度。

通过对上述研究分析得出，由于数据量较少、根据人工或通过传统统计方法所选取特征对模型精度有较大影响等因素，造成现有恐怖组织预测模型精度较低。为解决现有模型存在的问题，本文提出一种基于 GTD（全球恐怖主义数据库）数据及以特征工程作为特征选择方法的恐怖活动与恐怖集团关联性模型，支撑算法为随机森林及 KNN。通过关联性模型的构建进一步提升恐怖组织预测能力，以得到精度较高的预测结果。

2 恐怖活动与恐怖集团关联性模型

本文以 GTD（全球恐怖主义数据库）提供的 2000 年至 2018 年全球所发生的 121628 件恐怖袭击事件及其相关数据作为原始数据。GTD 主要记录全球 210 多个国家和地区所有能够统计到的恐怖袭击事件，信息包含了 1970 年以来所有有记载的恐怖袭击事件的时间、国家、地区、城市、袭击方式、袭击目标、采用武器、恐怖组织、动机和目的以及伤亡和财产损失情况等 135 个属性字段信息，是目前为止最权威最全面的恐怖袭击事件信息数据库，收集来源主要包括新闻等媒体报道和公开的纸质文件等。该数据库 1970 年至 1997 年间数据由全球情报侦探服务（ PGIS）收集，由于该部分数据为手稿形式，所以在 2001 年至 2005 年美国 START 员工将这些原始数据进行了数字化；1998 年 1 月至 2008 年 3 月间的数据和 2008 年 4 月至 2011 年 10 月的数据分别由恐怖主义情报研究中心（CETIS）和暴力组织研究协会（ISVG）收集；2011 年 11 月至今的数据由 START 员工进行收集。与此同时 GTD 的员工在 START 总部整合了 1970 年至 2018 年的全部数据，至此形成了现在我们看到的完整的全球恐怖主义数据库[14]。

根据数据库中恐怖主义袭击事件所包含的各项特征与恐怖组织特征之间相关性的分析为切入点，通过对数据库中凶手已知的恐怖袭击记录中的特征数据进行特征有效性分析，选取与恐怖组织具有较高关联系数的特征构建特征数据集，再对数据集中的数据进行 Min-Max Scaling、One-Hot Encoding 等处理后传入目前主流的机器学习分类算法当中进行模型训练，然后通过模型评价指标分析各个算法的优劣，最终构建出最优的恐怖活动与恐怖集团关联性模型，本文的模型构建流程如图 1 所示。

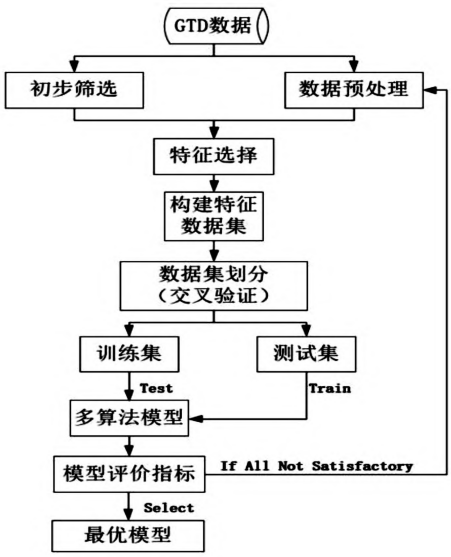

图1　模型构建流程图

2.1　**数据预处理**　在GTD所提供的数据中，“gname”这一列数据中记载了造成恐怖袭击事件的恐怖组织信息，对于凶手未知的事件，在该列数据中使用“Uknow”标记，根据此标记，首先提取出数据集中所有凶手已知的事件相关数据，再对数据进行去重，然后将其中的描述性文本特征进行删除，最后得到的59723条数据作为初筛数据。在得到初筛数据后，需将数据做进一步处理以得到达到符合模型输入条件的数据，具体处理流程如图2所示。

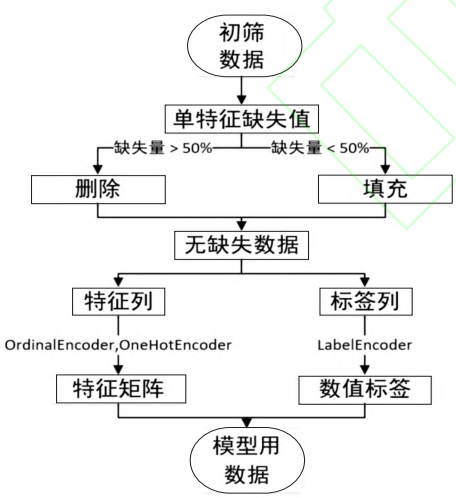

图2　数据处理流程图

首先将初筛中的缺失值进行处理，当某一特征变量中缺失值大于50%时认为该变量为无用变量直接删除即可。在删除无用特征变量后，对于剩余特征变量中存在的缺失值，通过探索性数据分析后根据变量特征采用众数填充，0值填充以及前项填充三种填充策略对缺失值进行相应的填充，最终得到无缺失特征集。

由于构造的特征集中各特征之间具有不同的量纲，且阈值范围差距较大，直接输入模型后会带来较大噪声，影响模型精度，因此需对特征集进行无量纲处理。

首先对连续型特征变量（如受伤总数、死亡总数等）进行MinMaxScaler处理，其原理是当数据x通过最小值Mean-Subtraction处理后，再按极差（Max-Min）缩放，数据移动了最小值个单位，最终处理后x^*的会被收敛到[0,1]的范围之间，公式如式1所示，归一化之后的数据服从正态分布。

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

其次，对于分类型特征变量（如攻击类型、武器类型等），由于其数据特点具有非偏序关系，所以对该类型数据进行独热编码；对于作为标签的gname列，首先通过Labele Encoding将其转化为分类变量，再对其进行独热编码。最终，经过归一化及独热编码处理后的特征集中所有特征变量均处于同一阈值范围内，达到模型输入要求。

2.2　**预测模型选取**　由于本文要解决问题为分类问题，故本文采用当前主流的随机森林、KNN、朴素贝叶斯三种分类算法做对比实验，运用k-fold cross validation（k折交叉验证）来验证模型的稳定性即模型的泛化能力，得到稳定模型后，通过模型评价指标Hamming Loss及0/1 Loss评价模型的效果及准确度最终确定最优模型。

随机森林（Random Forest）是解决分类问题最常用的算法，其原理是包含多个树模型的分类器，其优点在于可以通过输入多维特征，训练出高准确度的分类器，很好解决了单个决策树泛化能力弱的问题，且对于相对不平衡的数据集可以自动平衡误差。

区别于树模型根据判别类域分类，KNN（K-NearestNeighbor）是根据不同特征之间的距离值进行分类，即某一样本在特征空间中的k个最相邻的样本大多数属于同一个类别，其距离根据欧几里得度量（Euclidean Metric）来计算，其公式如式2所示（式中n表示特征维度，(x,y)表示特征点坐标）。其优点在于适合于多分类问题，且对稀有事件分类效果较好。

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_3)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

朴素贝叶斯分类器（Naive Bayes Classifier，NBC）是基于贝叶斯定理、采用“属性条件独立性假设”分类方法的分类器，即对已知类别，假设所有属性相互独立，每个属性独立地对分类结果发生影响[13]。由于NBC假设特征变量之间相互独立，所以当特征集数据类型不同时，分类性能不会受到太大影响，模型稳定性较高，其原理依据为：给定样本$x = (x_1, x_2, \cdots x_n)T$，其

中属于 ω 类的后验概率如式 3 所示(其中,n 是特征总数,x_k 是样本在第 k 个特征上的取值,$p(\omega)$ 为先验概率,可用 $P(\omega)=|D_i|/|D|$($|D_i|$ 训练集 D 的全部样本 $|D|$ 里属于 ω 类的样本数)求得)。

$$h_{nb}(x)=p(\omega_i|x)=\frac{p(x,\omega_i)}{p(x)}\propto p(x,\omega_i)=$$
$$p(\omega_i|x)p(\omega_i)=p(\omega_i)\prod_{k=1}^{d}p(x_k,\omega_i) \tag{3}$$

其中,n 是特征总数,x_k 是样本在第 k 个特征上的取值,$p(\omega)$ 为先验概率,可用 $P(\omega)=|D_i|/|D|$($|D_i|$ 训练集 D 的全部样本 $|D|$ 里属于 ω 类的样本数)求得。

k 折交叉验证是将数据集划分为 k 个均量的互斥子集,每次使用一个子集用作模型检验,其余 $k-1$ 个子集作为训练集,重复 k 次,使每个子集都得以验证后,平均 k 结果得到最终估值,本文采用 10 折交叉验证,其方法如图 3 所示。

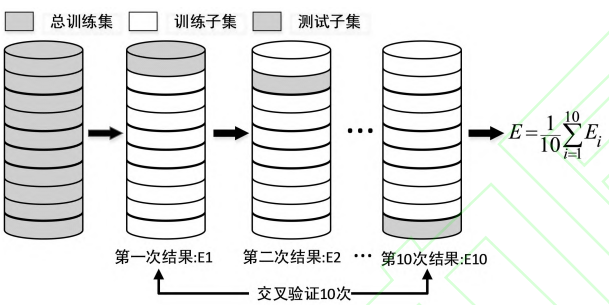

图 3　十折交叉验证

2.3　**关联性特征提取**　在建模过程中特征选择的好坏将直接影响模型的精度,选取与 label 相关度较高的特征能得到更为准确的预测结果,在数据预处理的基础上,区别于传统人为主观的选择特征,本文采用特征工程方法选择更贴切于模型,能使模型达到最好拟合效果的特征构建特征数据集。

对于恐怖活动所表现出的特征,刘亚男等提出恐怖主义的具体特征体现主要分为五类:袭击者、袭击目标(非战斗人员、设施等)、袭击手段(暴力、武力等)、袭击目的(社会性、政治性)、行为界定[15];王奇等提出恐怖活动特征指的是恐怖行为在计划实施的阶段所暴露出的有迹可循、有章可依的线索和特征,主要包括时间特征、空间特征、行为表现特征、目标对象特征四个方面[16];刘云虹等依据恐怖活动事件的性质、袭击发生地域、事件的大小规模、人数的伤亡情况和财产的损失情况五个方面选择了共 GTD 数据库中的 24 个特征作为特征关联的基础指标[17-21]。

本文在综合现有文献所选择的特征的基础上再将数据库中的所有数值变量及分类变量数据都加入预选特征中,构建整体特征数据集,给算法以更多的特征选择范围,根据多次迭代选择最优特征。

对得到的整体特征数据集,进行关联性特征提取。本文采用方差过滤、相关性过滤中的卡方过滤及 Embedded 三种特征选择方法相结合选择模型所需要的特征。

方差过滤是依据统计检验中的相关指数通过特征本身的方差来对特征进行筛选,所以用来初步筛选,优先消除方差为零的特征变量。

在初筛后为防止与标签无关的特征对模型带来噪音,需再对其进行相关性过滤,其中卡方过滤是相关性过滤中专用于离散标签的过滤方法,其原理是计算各非负特征与标签之间的卡方统计量并对其进行由高到低排名,最后根据某个评分指标选出指标较高的前 n 个特征;由于卡方过滤需要输入非负值所以前文在数据预处理阶段对特征变量采用了归一化处理。为了获得最佳的模型效果,为了获得最佳的模型效果,我们需要通过学习曲线来确定选择多少个特征,即确定 n 的值,图 4 为卡方过滤学习曲线。

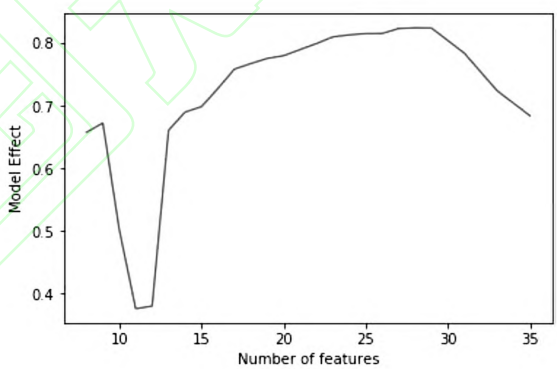

图 4　卡方过滤学习曲线

通过学习曲线可以看出当选择的特征量在 27-30 之间时模型效果达到峰值,当超过 30 个特征后模型出现过拟合现象,即根据卡方统计量由高到低选取前 27~29 个特征变量能使模型达到最好的拟合效果。

为进一步验证卡方过滤结果的准确性,本文引用 Embedded(嵌入法)对特征进行二次检验。

嵌入法与前者不同之处在于嵌入法是特征变量的选择与算法的训练是同步进行的,其选择特征的准则是通过模型的训练得到各个特征变量对于模型贡献度及重要性的权值系数,再由系数由大到小选择特征,该方法的优点在于特征变量选择更加贴切模型效果本身。嵌入法特征选取流程如图 5 所示。

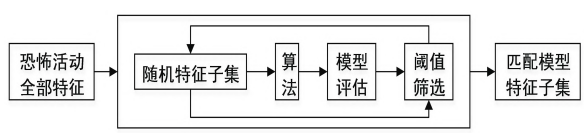

图 5　嵌入法选取流程

本文选用随机森林中的 feature_importances_属性

调取各特征的重要性阈值,用阈值作为权重系数根据阈值选择特征变量的数目,同样的,通过学习曲线来选择让模型达到最优效果的阈值,嵌入法学习曲线如图6所示。

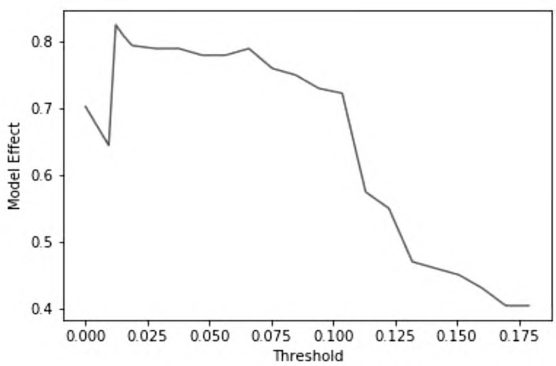

图6 嵌入法学习曲线

通过学习曲线得知阈值在介于0~0.025之间所选取的特征数能使模型效果达到峰值,通过二次细分,得到所选特征数为29个特征变量,与卡方过滤所得27~30个特征数量吻合。因此在本文构建的恐怖活动与恐怖组织关联性模型中选取通过嵌入法得到的29个特征变量作为最终的特征数据,所选的29个特征变量及数据集中对应变量名为:年(iyear)、月(imonth)、日(iday)、凶手数量(nperps)、死亡总数(nkill)、凶手受伤人数(nwoundte)、人质/绑架受害者总数(nhostkid)、入选标准1(crit1)、入选标准2(crit2)、入选标准3(crit3)、是否为持续事件(extended)、事件组的一部分(multiple)、国家(country)、地区(region)、成功的攻击(success)、自杀式袭击(suicide)、武器类型(weaptype1)、武器子类型(weapsubtype1)、目标/受害者类型(targtype1)、第二目标/受害者类型(targtype2)、目标/受害者的国籍(natlty1)、第二目标/受害者的国籍(natlty2)地理编码特征(specificity)、攻击类型(attacktype1)、第二攻击类型(attacktype2)、索要赎金(ransom)、财产损害程度(propextent)、绑架/人质结果(hostkidoutcome)、国际杂类(INT_MISC)。

3 模型评价指标

由于本文所建模型解决的为多分类问题,所以选取Hamming Loss及0/1 Loss作为模型的评价指标。

Hamming Loss公式如式4所示:

$$H_{Loss} = \frac{1}{N}\sum_{i=1}^{n}\frac{XOR(x_{ij},y_{ij})}{L} \tag{4}$$

其中n为样本总数,L为标签总数,x_{ij}表示在第i个模型预测结果中j分量的真值,y_{ij}表示第i个结果中j的预测值,XOR表示异或运算。H_{Loss}的取值范围为[0,1],其值越趋近于0,模型效果越好。

0/1 Loss公式如式5所示:

$$0/1_{loss} = 1 - \frac{1}{N}\sum_{i=1}^{n}\delta(C_i^{'},C_i) \tag{5}$$

其中C_i为数据集中第i个样本的真值,$C_i^{'}$表示模型给出的预测值,n为样本总数,其中当且仅当$C_i^{'} = C_i$时$\delta(C_i^{'},C_i) = 1$。0/1 Loss表示了第i个实例模型预测值$C_i^{'}$与真值C_i不同的样本个数占全部样本的比重,与Hamming Loss相同,取值范围为[0,1],值越趋近于0,模型效果越好[23]。

4 实验结果分析及改进

在进行模型训练前,首先对数据集中出现次数少于三次的恐怖组织样本进行删除,因为将样本量过少的数据输入模型后会对产生噪声影响模型精度。在对样本量过少的数据删除后,将经过数据预处理及特征筛选后的特征数据集导入模型进行训练,对三种算法通过10折交叉验证得各算法两种评价指标,将指标值进行百分转化的对比结果如图7所示。

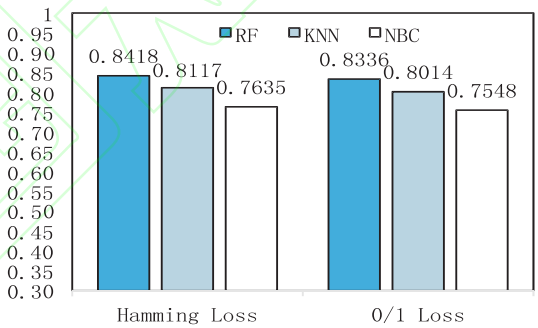

图7 模型效果对比图

由柱状图中数据可知,RF模型效果在两种评价指标中评分均为最好,KNN次之,NBC表现效果不理想,分析原因是由于朴素贝叶斯假设各特征间相互独立,但现实中本数据集间各特征变量间有一定的相关系数,所以该模型效果不理想。对于KNN模型,由于模型原理需将所有特征都计算欧几里得度量,所以在运行时间上远大以RF及NBC。无论时间及准确度,RF均优于其他两种算法,通过精度(Precision)计算随机森林模型精确度,其计算公式如式6所示,公式中TP、FP含义如图8混淆矩阵所示,得到基于随机森林构建的模型预测精度为83.24%。

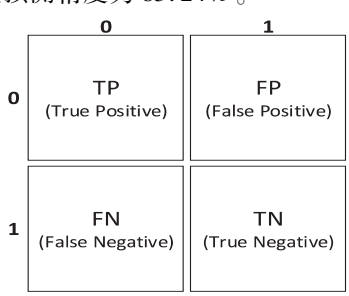

图8 混淆矩阵

$$p = \frac{TP}{TP + FP} \tag{6}$$

在对模型使用的数据集进行统计分析可得,在模型使用的58614条样数据中,出现的恐怖组织个数为688个,其中每个恐怖组织所包含样本数最少为三起最多为8723起,样本差异明显,为样本不均衡数据集;为进一步提升模型精度,特别是针对预测模型在少数类样本上的精度,在前期处理的基础上需对恐怖组织在样本量维度上进行进一步细分。通过对恐怖组织样本量进行聚类分析后,根据恐怖组织出现频率及其样本量进行样本分段,分为高频、中频、低频三段,分段如表1所示。

表1 恐怖组织频率分段

出现频率	袭击次数区间	样本个数	恐怖组织个数
高频	[100,8723]	48677	65
中频	[21,99]	6542	147
低频	[3,20]	3395	475

将经过频率分段处理后的数据分别输入本文所采用的三种算法模型再次进行训练,对训练好的模型进行精度对比,其结果如图9所示。

由图9中信息可知,使用分段后数据训练的模型三者精度均有明显提升,特别的,由于KNN模型在少数类样本上具有较好分类预测效果,在对低频恐怖组织进行预测时KNN模型精度可达到86.14%。

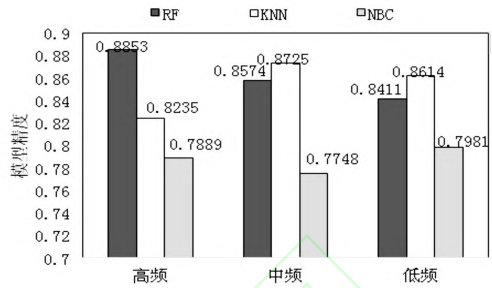

图9 频率分段模型精度对比图

因此,在对凶手未知的恐怖活动嫌疑人进行预测时,可分别使用随机森林模型在高频段恐怖组织中得到预测结果,使用KNN模型在中频及低频段得到预测结果,在得到的三个预测结果中锁定恐怖活动嫌疑人。

为检验模型效果,本文基于蓄水池抽样算法,从GTD数据库中抽样选取2007-2018年之间的10条凶手已知的样本数据,蓄水池抽样算法保证了每个样本被抽中的概率均相等,将这十条样本数据代入预测模型推断其嫌疑人,并与实际造成该事件的恐怖组织进行对比,其结果如表2所示。

表2 抽样预测结果对比表

恐怖袭击事件编号	实际凶手	预测结果1	预测结果2	预测结果3
201203090020	Popular Resistance Committees	Palestinian Islamic Jihad (PIJ)	Popular Resistance Committees	Iraq's Jihadist Leagues
201509140072	Hindu Morcha Nepal	Houthi extremists	Lashkar-e-Jhangvi	Hindu Morcha Nepal
200805170010	New People's Army	New People's Army	Albanian extremists	Right Sector
200805140021	Conqueror Army	Chechen Rebels	Muslim Brotherhood	The Revolt
201511150043	Anti-Muslim extremists	People's War Group	Anti-Muslim extremists	Weichan Auka Mapu
200904160038	Al-Shabaab	Al-Shabaab	Oglaigh na hEireann	Nepal Defense Army
201111080021	Militants	Fulani extremists	Militants	Ijaw extremists
201005100035	Al-Qaida in Iraq	Al-Qaida in Iraq	Sunni Muslim extremists	Tuareg extremists
201609040087	Hasam Movement	Maoists	Hizb-I-Islami	Hasam Movement
200904190003	Taliban	Taliban	Taliban (Pakistan)	Ukrainian nationalists

由表中数据可得,在抽取的10条样本数据中,从模型预测出的三个嫌疑人中找到实际造成该事件的恐怖组织的概率为90%;其中有一条数据预测结果出错,分析原因得知,该恐怖组织在GTD数据库中只存在两条样本数据,由于其样本量过少,在模型中无法对其进行分类训练,故无法得出准确推断。总体而言,通过本文所构建模型在不同频率范围得出的融合结果整体精度较高,为可行性预测模型。

5 结 语

本文将特征工程应用到恐怖组织预测模型中的特征选取上,使得所选特征动态契合不同算法,使模型得到更好的拟合效果及泛化能力,大幅度提升模型精确度;并且针对不同恐怖组织所包含样本量的样本不平衡问题,采用根据样本频率分段预测的方法,结合随机森林模型及KNN模型在不同样本量的特点,进一步提升了恐怖组织预测模型性能,特别是针对低频恐怖组织的预测性能得到明显提升。通过本文所构建的恐怖组织预测模型可大范围缩小恐怖活动嫌疑人排查范围,进一步对恐怖分子做出针对性的打击。

由于低频恐怖组织所包含样本数据较少,使得预测模型在低频恐怖组织中性能未到达理想效果,在今后研究中将进一步对低频恐怖组织预测进行深入研究,提升其预测精度。

参 考 文 献

[1] 舒洪水,李燕飙. 基于大数据视角的恐怖袭击特点与趋势分析——以 GTD 中的 7133 次恐怖袭击为样本[J]. 情报杂志,2019,38(11):87-93,198.

[2] 纳兰星舟,毛欣娟. “独狼”式恐怖袭击的主要特征及防控思考[J]. 武警学院学报,2019,35(6):68-72.

[3] Terrorism; New findings from russian academy of sciences in the area of terrorism published (myths and reality of "lone wolf" terrorism in the context of islamist extremism)[J]. Bioterrorism Week,2020.

[4] Jude McCulloch , Jasmine McGowan. Lone wolf terrorism through a gendered lens: Men turning violent or violent men behaving violently? [J]. Springer Netherlands,2019,27(3).

[5] 强 琪,任延涛,刘 通. 对“独狼”恐怖分子行为动因分析及防控思考[J]. 辽宁公安司法管理干部学院学报,2019(6):21-26.

[6] 白海娟,屈耀伦. 新疆恐怖主义犯罪的特点和新动向[J]. 南都学坛,2016,36(6):75-79.

[7] 郝蒙蒙,陈 帅,江东. 中南半岛恐怖袭击事件时空演变特征分析[J]. 科技导报,2018,36(3):62-69.

[8] Xue A , Wang W , Zhang M . Terrorist organization behavior prediction algorithm based on context subspace[M]. Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2011.

[9] Zou B, Nurudeen M, Zhu C. A neuro-fuzzy crime prediction model based on video analysis[J]. ChineseJournal of Electronics, 2018,27(5):968-975.

[10] Li Z, Sun D, Li B. Terrorist group behavior prediction by wavelet transform-based pattern recognition[J]. Discrete Dynamics in Nature and Society, 2018,2018:1-16{17}

[11] 战 兵,韩 锐. 基于隐马尔可夫的恐怖事件预测模型[J]. 解放军理工大学学报(自然科学版),2015,16(4):386-393.

[12] 李 慧,张南南,曹卓,等. 基于机器学习的恐怖分子预测方法研究[J/OL]. 计算机工程:1-7[2020-04-02]. https://doi.org/10.19678/j.issn.1000-3428.0053521.

[13] 王行甫,杜 婷. 基于属性选择的改进加权朴素贝叶斯分类算法[J]. 计算机系统应用. 2015,24(8):149-154.

[14] 彭如香,张奥博,杨 涛,等. 基于 GTD 的全球恐怖主义活动现状与发展趋势研究[J]. 计算机应用与软件,2019,36(1):1-5+21.

[15] 刘亚男,任 婧,邓小勇,等. 恐怖活动犯罪构成特征分析[J]. 法制与经济,2017(2):111-113.

[16] 王 奇,田一鸣. 全球恐怖活动的 GTD 数据分析与我国应对之策[J]. 犯罪研究,2018(2):87-96.

[17] 马天成. 恐怖活动犯罪类型化研究[D]. 中国人民公安大学,2019.

[18] 刘云虹. 基于恐怖活动特征分析的反恐情报预警模式构建[D]. 中国人民公安大学,2019.

[19] 王向爱,庄元强,谢为顿,等. 对恐怖袭击事件记录数据的量化分析与研究[J]. 经济数学,2019,36(3):95-103.

[20] 周松青. 全球恐怖主义数据库及对中国反恐数据库建设的启示[J]. 情报杂志,2016,35(9):6-11.

[21] 叶小琴,康倩飞. 我国暴恐犯罪的特点与预防:基于 GTD 数据库的统计分析[J]. 犯罪研究. 2018(1):18-27.

[22] 连芷萱,夏一雪,史路遥,等. 基于 logistic-ABC 的恐怖活动风险因素识别与指标评级模型研究—俄罗斯 2006-2016 年恐怖活动的实证分析[J]. 情报杂志. 2018,37(11):23-30.

[23] 李 锋. 基于标签特征和相关性的多标签分类研究[D]. 西安电子科技大学,2019.

[24] 韦 灵,黎伟强. 基于机器学习的中文文本自动分类的实践研究[J]. 智库时代,2019(46):265-266.

[25] 朱柯睿,周瑞鑫,康世举,等. 两种监督机器学习算法在 Fermi BCU 分类评估中的应用[J]. 云南师范大学学报(自然科学版),2019,39(5):1-5

[26] Zhang. Cost-sensitive KNN classification[J]. Elsevier, 2019 (prepublish).

[27] Xianglong Luo, Danyang Li, Yu Yang. Spatiotemporal traffic flow prediction with KNN and LSTM[J]. Hindawi,2019,2019.

[28] Ma Zong fang,Tian Hong peng,Liu Ze chao. A new incomplete pattern belief classification method with multiple estimations based on KNN[J]. Elsevier B. V. ,2019,90.

[29] 肖跃雷,张云娇. 基于特征选择和超参数优化的恐怖袭击组织预测方法[J/OL]. 计算机应用:1-9[2020-07-06]. http://kns.cnki.net/kcms/detail/51.1307.TP.20200424.1203.004.html.

[30] 邱凌峰,韩昕格,胡啸峰. 基于机器学习的恐怖袭击事件后果预测方法研究[J]. 中国安全生产科学技术,2020,16(1):175-181.

[31] 王莲花,李筱烨,高 楠. 基于灰色理论的恐怖袭击事件预测与分析[J]. 数学的实践与认识,2019,49(12):174-182.

[32] 姜立宝,陈昱帆,俞 璐,等. 一种基于聚类的反恐数据分析方法[J]. 情报探索,2019(6):74-77.