Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach

摘要：

恐怖事件可能对整个社会造成深远影响。发现恐怖袭击的规则对全球反恐战略具有重要意义。在本研究中，我们展示了一种新方法，该方法使用相对流行且强大的机器学习方法，基于多种资源，长时间序列和全球分布的数据集来模拟全球范围内的恐怖袭击风险。采用1970年至2015年的历史数据来训练和评估机器学习模型。该模型在预测2015年可能发生恐怖事件的地方方面表现相当不错，成功率为96.6％。此外，值得注意的是，具有优化调整参数值的模型成功地预测了2,037个恐怖袭击之前从未发生过的恐怖主义事件。

介绍：

恐怖主义被认为是对社会的主要威胁[1-3]。根据全球恐怖主义数据库（GTD），仅在2015年全球就发生了14800多起不同类型的恐怖事件[4]，这不仅造成38430人死亡，而且还引起世界恐慌[5-7] 。已经做出巨大努力来寻求与恐怖主义威胁相关的各种问题的解释[8,9]。然而，对某一事件发生的预测仍然是一项极其复杂和不确定的困难任务[10,11]。目前，数据驱动模型是大量高维训练数据的直接学习模式，在各个领域都取得了一些成功[12-15]，为解决这一问题提供了新的视角。

以前的研究表明，自然[16]和社会[17]因素被用于基于代理的方法来模拟暴力冲突的潜在未来情景，但潜在的机制仍然没有被充分理解。一切都是连通的。也许，恐怖袭击有一定的规律性[18]。因此，地理学家的典型直觉提出了一个问题，即恐怖事件的发生是否可能受到地理，自然和社会问题（因素）的影响或控制。如果存在这些模式，则可以通过机器学习方法来发现它们。

鉴于全球数据的可用性，我们收集了三个社会因素，如种族多样性，主要毒品区域和人口密度。此外，还包括夜间照明，这已被广泛验证用于监测经济活动[19]。同时，在本研究中选择了一些天然元素，包括平均降水量，平均温度和地形。考虑到这些因素在空间，纬度，经度，到主要通航湖泊的距离，到无冰海洋的距离以及到主要可航行河流的距离之间的差异被用来表达地理空间信息。结果，在全球范围内（0.1×0.1度）量化了总共12个因子。由于马里兰大学STAR团队的出色工作，GTD成立，详细记录了超过150,000个活动。我们对1970年至2014年发生的恐怖事件的位置进行了栅格化，根据地点的频率和后果（即死亡人数）将单位划分为2类。

具体流程图如图1所示。为了创建仿真模型，我们开发了一系列程序，并使用地理信息系统（GIS）软件预处理各种类型的数据并选择样本数据集（参见方法）。在本研究中使用了三种相对流行的机器学习算法：神经网络（NNET），支持向量机（SVM）和随机森林（RF）。此外，方法中描述了机器学习模型的训练方法和测试。值得注意的是，研究范围仅限于75̊N和55̊S之间的区域。

方法：

本研究中使用的数据类型和来源如表1所示.GIS软件和C ++编程语言用于预处理数据集，包括ArcMap 10.2（http://www.esrichina.com.cn/），GDAL 2.1 .0（http://www.gdal.org/）和Proj.4（https：// github.com/OSGeo/proj.4）。选择WGS-84作为本研究中使用的所有数据的地理坐标系。

基于GTD，我们可以获取每个事件的位置，并将1970年至2014年世界各地的恐怖事件信息转换为栅格数据。选择0.1×0.1度分辨率的像素作为统计上确定恐怖事件数量和死亡总数的单位。在本研究中，如果过去发生过造成人员伤亡的恐怖袭击，评估单位被认为具有较高的恐怖袭击概率;否则，概率很低。给出了这些类型的类值1和0。另外，我们可以获得每个单元中心点的纬度和经度信息。

基于G-经济学4.0，在世界经济活动的数据集，我们可以得到的5个因素，即主要通航的湖泊（公里），距离距离主要通航河流（公里），距离光栅数据不冻海洋（公里），平均降水量（毫米/年）和平均透射电镜perature（C）从1980年到2008年。然后，我们通过使用ArcMap 10.2因为每个0.1度纬度分辨率重新采样上述光栅数据到0.1度的经度在全球范围内通过1度纬度分辨率在1度经度下测量G-Econ的地面观测。

种族分布图是从GeoEPR以0.01度经度通过使用ArcMap的10.2“多边形到栅格”功能0.01度纬度的分辨率的。在本研究中，我们假设种族群体在恐怖事件方面没有内在差异，但不同文化信仰与邻近地区的冲突可能导致恐怖事件。我们使用0.1×0.1度分辨率像素作为单位来统计确定每个地区不同种族群体的数量。 主要毒品区域来自世界药物报告和国家行政边界（http://www.gadm.org/），使用ArcMap 10.2“Polygon to Raster”和“Resam-”功能。其余因素（夜间灯光，人口密度和地形）从各自的数据源计算，并使用ArcMap 10.2重新采样为0.1×0.1度分辨率像素。

另一点需要注意的是，上面提到的十二个潜在因素具有不同的单位。为了促进模型的训练步骤，将它们标准化为下面的等式1：

其中Pj0是0到1之间的归一化值; Pj是因子的价值; Pmin是最小值; Pmax是最大值; n是数据的数量。

数据采样：

从每个数据层的相同位置收集单元属性的值，并一起构成样本行。如果某些恐怖袭击的地点很接近，它们将被记录在同一统计单位。此外，排除了缺少某些因素信息的小部分单位。在目前的研究中，选择恐怖袭击导致伤亡人数超过0的像素，并从剩余像素中随机选择相同数量的像素（图2）。这一程序确保了导致伤亡的恐怖袭击样本数量只占整个样本的一半。最后，我们可以获得15970个样本数据。为了训练和测试机器学习模型的性能，随机选择四分之三的样本数据作为训练数据，剩余的数据用作测试数据，几乎均匀地分布在全世界。

学习模型：

我们使用了64位版本的R版本3.30，这是一种用于统计计算的开源语言和软件[20]，用于构建模型，调整参数和执行精度评估。本研究采用了三种相对流行且强大的机器学习算法（NNET，SVM和RF）。 “插入符号”包[21]和内置函数用于训练每个模型并在单个一致的环境中调整其相关参数。

对于NNET，我们使用反向传播神经网络（BPNN）算法，该算法具有强大的突触修改规则，允许任意连接的神经网络开发适合特定任务域的内部结构[22]。对BPNN的研究表明，增加隐藏神经元的数量可以提高分类准确度[23,24]，大的衰减率会导致非收敛[22]。因此，当在隐藏层和输出层中运行具有S形激活函数的反向传播神经网络时，调整衰减值和隐藏层的隐藏单元的数量。

对于基于SVM的仿真，径向基函数（RBF）内核被考虑用于本研究。先前对SVM算法的研究强调，RBF内核中使用的两个参数（cost，sigma）对分类精度有影响[25]。大成本意味着对错误赋予更高的惩罚[26]，而增加的西格玛则影响分离超平面的形状[27]。因此，在使用RBF内核进行仿真时，我们需要微调成本和sigma参数。

对于使用RF模型构建的模拟，使用默认树数（500），因为大于500的值无法显着提高RF算法的性能[28]。 mtry参数控制随机抽样的变量数量，作为每个分割的候选者，并且已知会对整体分类准确性产生影响[29]。对于二进制分区，可以使用少量随机选择的变量，例如mtry参数的值，它等于数据集中变量数的平方根[30]。在本研究中，运行RF算法时仅调整了mtry参数。

结果：

在使用机器学习模型来模拟恐怖袭击的风险之前，我们需要调整一些参数。曲线下面积（AUC）用作衡量模型性能的指标。较大的AUC值表示较高的准确度。基于此标准，我们选择了具有优化调整参数值的模型。对于NNET模型，获得了具有30个隐藏单位和0.1的衰减值的最高AUC值（0.966）（图3A）。对于使用SVM模型的模拟，检查了成本参数（2,4,8,16和32）的总共5个值和sigma参数的6个值（0.08,0.09,0.1,0.12,0.14和0.16）。基于获得的最高AUC（0.962），为SVM模型选择成本值32，而将西格玛参数的值设置为0.09（图3B）。基于RF模型检查总共12个mtry参数值（1,2,3,4,5,6,7,8,9,10,11和12）用于模拟。对于RF模型，mtry值为2获得最高AUC值（0.974）（图3C）。可视化重采样分布，在训练样本的交叉验证过程中，RF模型比NNET和SVM模型具有更好的性能。此外，优化模型的AUC值超过0.95，证明恐怖袭击与12个因素之间存在一些模式（图3D）。

基于应用于训练数据集的机器学习模型之间的比较，我们发现RF模型获得最高的AUC（1）值，接着是NNET（0.980）和SVM（0.976）模型（图3E）。对于应用于验证数据集的机器学习模型，观察到相同的总趋势; RF模型具有最高的AUC（0.976）值，其次是NNET（0.971）和SVM（0.966）模型（图3F）。此外，McNemar [31]测试用于评估使用验证样本时不同机器学习模型之间是否存在统计学上显着的差异。 McNemar测试显示，基于验证数据的机器学习模型之间的差异没有显着差异（p> 0.05）。

根据比较分析的结果，我们利用获得最佳性能的RF模型来预测全球所有地区的恐怖袭击风险。结果表明，恐怖袭击的高风险区域主要集中在南亚，中东，中非，北非，南美洲西北部，南美洲南部，南欧和西欧（图4A）。

我们将2015年发生的恐怖事件的验证结果分为4类（A，B，C，D），并强调在这一步骤中排除了在缺少某些因素信息的单位中发生的一些恐怖袭击。一般来说，高危地区发生了11,632起恐怖袭击事件，射频模型的预测准确率为96.6％。更重要的是，RF模型成功预测了2037个恐怖主义事件发生地点，其中1970年至2014年期间没有发生恐怖袭击但发生在2015年（图4B）。除此之外，最佳模型在预测2013年和2014年可能发生恐怖事件的地方方面表现相当不错，成功率分别为96.0％和94.7％。

此外，我们也有兴趣了解12个因素与恐怖袭击之间的机制。因此，我们试图统计分析每个因素的歧视力。使用“Boruta”包[32]进行测量，使用置信水平的默认值（0.01），并将重要源运行的最大数量设置为100.图5显示上述地理，自然和社会因素对于模拟全球地区恐怖袭击的风险至关重要。此外，结果表明，人口密度，夜间灯光和单位的位置（纬度，经度）比区别像素是否属于高风险单位的剩余因素发挥更重要的作用，这类似于从F-score方法。人口密度因子具有最高的平均重要性（66.49），其次是纬度（64.45），经度（62.50），夜间灯光（61.13），到主要通航河流的距离（58.88），到主要通航湖泊的距离（53.28）和主要毒品区（52.86）。鉴于主要毒品区域的平均重要性，参考文献8可能是合理的，可以解释的，该文章写道恐怖主义团体将毒品交易作为资金来源征税。

结论：

在这项工作中，我们展示了一种基于大量数据的新型战略，该战略使用相对典型且稳健的机器学习模型，虽然不是最先进的，但可以模拟全球所有地区的恐怖袭击风险，这被证明是合理的。值得注意的是，使用机器学习模型成功预测恐怖袭击的风险主要取决于大量的高维数据，包括地理，自然和社会因素，这些数据使用数据集预处理量化为空间（参见方法） ）。此外，我们认识到其他研究人员对数据源的贡献（表1）。这些数据可能不是全球恐怖袭击模拟风险中最具代表性的数据，但可以免费下载并量化到该空间。此外，基于Boruta算法总结了各因素的重要性，其结果是数据驱动的，具体原因需要进一步研究。

不幸的是，在这项研究中发生了一些恐怖事件，例如尼斯袭击，布鲁塞尔炸弹袭击，奥兰多夜总会袭击，维尔茨堡火车袭击以及慕尼黑枪击事件，造成大量人员伤亡。为了进一步验证预测的正确性，我们根据预测结果测试了这些恐怖事件的位置，这表明上述恐怖袭击发生在高风险地区。此外，基于验证样本，机器学习模型（NNET，SVM和RF）之间没有发现显着差异。

下一步将侧重于更精细的时空尺度，以分析恐怖事件，例如不同地区恐怖袭击的分布以及不同年代之间恐怖袭击的特征。此外，当使用具有丰富数据的机器学习模型时，我们将更好地利用来自GTD的信息，而不是仅使用每个事件的位置和伤亡人数，这可能揭示隐藏在数据中的更多关键原则。