数据驱动的结构化交通风险预测模型

陈光晓，李　营，王涛

（中国航空工业集团公司济南特种结构研究所 信息档案技术研究室， 山东 济南 250100；）

摘　要：交通事故救援的时效性和有效性对于保障公民生命财产安全至关重要，如果能够对交通事故风险进行准确预测，并在识别到潜在风险后及时进行现场勘测与救援，将显著提升救援响应的时效性，有效维护公众安全。现有的交通风险预测模型，一方面受分析人员对诱发交通事故发生因素的考虑不足，以及经验或知识的局限性的影响，无法准确预测；另一方面，模型大多只通过评判交通事故发生概率来进行风险输出，而忽略了事故严重程度对风险值的影响。设计了一种结构化的交通风险实时性预测模型，将交通风险分解为事故发生概率与事故严重程度两个层次结构，更有助于理解和分析不同因素在整体风险中的角色，在应对动态交通环境的挑战时更具优势。首先对数据对象进行了必要的数据预处理，然后通过自相关分析、Person 相关性分析等数据分析技术，对影响交通事故的各种时空因素进行数据挖掘和数据深入分析，并根据分析结果完成特征的选择，最后通过 PU-Learning 算法结合随机森林模型，实现交通风险预测模型的构建。经实验分析，与基准模型相比，该模型具有更良好的预测性能。

关键词：数据挖掘；交通风险预测；机器学习

0 引言

随着人民生活水平的普遍提升，对汽车的需求量逐年增加，随即的交通事故数量也不断增长，据世界卫生组织数据显示，全球每年有超过 130 万人在道路交通事故中丧生，每 24 秒就有一人遇难，是年轻人和儿童的主要死因，同时还有 2000 至 5000 万人受到严重身体伤害，给个人、家庭和整个国家造成巨大的经济损失。如何通过新兴技术对交通事故进行有效预测和及时响应，将交通安全隐患消灭在萌芽阶段，提升道路交通事故预防能力，减少这种悲剧的发生，已成为国内外政府和科研机构的热点研究之一。

道路交通是一个时变性显著、关联性较强的复杂系统，造成道路事故的因素有很多，包括空间相关性、时间动态交互作用和外部影响等多个因素。如果能利用历史数据和数据统计，对道路交通事故与动态影响因素的内在联系进行探究，就能对潜在的交通事故风险进行科学估计与预测，可以做到在事故发生前通过主动采取相应的交通安全管理措施，对事故风险进行快速响应，从而降低事故发生的风险概率或严重程度，因此交通事故风险预测本质上是一个时空数据挖掘（Spatio-Temporal Data Mining，STDM）问题。

1 数据分析与特征选择

1.1 数据集描述

数据选自 Moosavi 等人在 kaggle 中发布的美国交通事故数据集（5.0 版本），其中包含了 2016 年 2 月至 2020 年 6 月美国 49 个州的全国范围交通事故数据，约有 150 万条事故记录，每个事故记录包含 47 个描述交通事故的属性。由于本研究主要针对城市级进行交通风险预测，从各城市规模、交通事故数据量、人口密度、交通出行需求及城市影响力多个指标综合考虑，选取城市洛杉矶作为主要的研究区域，数据集主要属性及其描述如表 1 所示：

表 1　主要属性及描述

Table 1　Main attributes and descriptions

	属性			描述
严重程度	Severity			事故严重程度（值从 1 到 4）
发生时间	Start_Time	End_Time		事故开始时间、结束时间
位置信息	Start_Lat	Start_Lng	State	事故 GPS 坐标、所属州和城市、所在街道、最近机场气象站
	City	Street	Airport_Code	
天气条件	Temperature	Humidity	Wind_Direction	事故发生地温度、湿度、风向、风速、大气压、能见度等
	Wind_Speed	Pressure	Visibility	
POI 标志	Junction	Traffic_Signal	Crossing	事故发生地是否有交叉口、交通信号灯、减速带等标志
	Bump	Amenity	Railway	

1.2 数据预处理

1.2.1 缺失值分析和处理

图 1 可视化了数据集的数据缺省情况，可以看到有 19 列都存在数据缺失，其中“Number”、“Wind_Chill(F)”和“Precipitation(in)”三列缺省最为严重，分别有的 29.3%、33.8% 和 69.1% 的缺失。

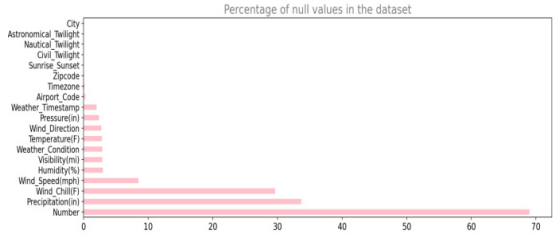


图 1 数据集缺失情况

Fig.1 Missing condition of the data set

（1）数据删除

缺失值非常大的这三列，与交通风险预测的关联性不大，直接删除。此外，与总体样本相比，“City”、“Timezone”等特征属性中缺失值非常少，这里也直接采用数据删除的方式对其处理。

（2）数据插补

对于“Pressure(in)”、“Wind_Direction”、“Temperature(F)”、“Visibility(mi)”、“Humidity(%)”和“Wind_Speed(mph)”等缺省很小的天气特征列，采用数据插补的方法来处理缺失项，来保留其中隐含的有价值的信息，处理过程如下：

1）选择位置特征“Airport_Code”和时间特征“Start_Month”，按位置和时间对天气特征进行分组。

2）对“Wind_Direction”这种离散型天气特征，将其缺失值将替换为每组的众数，而对于其他连续型天气特征，则用中位数填充。

1.2.2 数据变换

（一）天气数据简化

为使得模型更容易掌握和记忆不同天气类型的特征和变化规律，按表 2 所示的包含关系，将相似天气类型进行合并简化为 7 项：

表 2 天气数据简化

Table 2 Simplified weather data

Weather_Condition	Values
Clear	Clear
Cloud	Cloudy, Funnel Cloud, Scattered Clouds, Overcast
Rain	Light Rain, Rain,T-Storm, Light Thunderstorm, Light Thunderstorms, Thunderstorm, Thunderstorms,
Heavy_Rain	Rain Shower, Rain Showers, Light Rain Shower, Light Rain Showers, Heavy Rain, Heavy Rain Showers, Heavy T-Storm
Snow	Snow, Light Snow, Light Snow Grains, Snow Grains,Low Drifting Snow, Ice Pellets, Light Ice Pellets, Sleet, Light Sleet,
Heavy_Snow	Heavy Snow, Heavy Sleet, Heavy Ice Pellets, Light Snow Shower, Snow Showers, Squalls
Fog	Fog, Light Fog, Partial Fog, Patches of Fog, Shallow Fog

（二）风向数据简化

同样，按照表 3，对风向数据进行简化处理：

表 3 风向数据简化

Table 3 Simplification of wind direction data

Wind_Direction	Values
E	E, East, ESE, ENE
W	W, West, WSW, WNW
S	S, South, SSW, SSE
N	N, North, NNW, NNE
NE	NE
SW	SW
NW	NW
CALM	Calm
VAR	Var, Variable

1.2.3 基于聚类的处理

虽然 POI、街道或者人口在一定程度上可以体现空间变化，但城市内不同地点（即市中心核心区和住宅区）的交通事故模式（事故起因、类型、交通执法）仍然存在很大差异。为此，在建模预测之前，首先通过 K-Means 聚类方法，对数据中的空间异质性进行了处理，即根据数据内部的相似性，将样本集划分为多个类别。

K-Means 算法通过距离来度量样本间的相似程度，然后将相似的样本划进同一个簇。实验显示 k=7 时运行结果最优，因此这里设置簇的个数为 7，图 2 为对 2020 年发生在洛杉矶的交通事故进行聚类的结果，七个不同颜色的区域代表了不同相似程度的类别，红色圆点指示了编号为 1-7 的七个聚簇的中心，然后根据每个样本所属的聚簇中心编号，添加一列新的属性“cluster_id”。

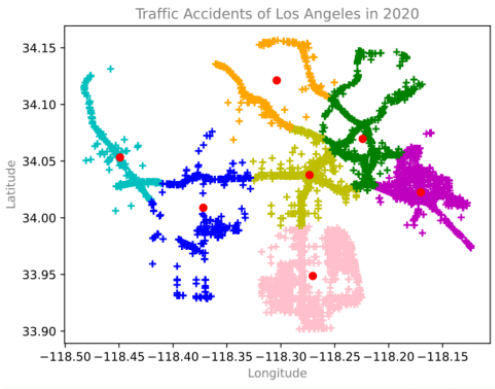


图 2 基于 K-Means 的相似程度划分

Fig.2 Division of the degree of similarity based on K-Means

1.3 数据分析与特征选择

1.3.1 特征重要程度分析

图 3 显示了通过极端随机树（Extremely Randomized Trees，ERT）集成学习模型计算的得到的前 22 个重要的特征。可以看出，在交通事故严重程度模型中，最重要的预测特征是小时（Hour）、周（Weekday）和年份（Year），其他重要的预测因子包括分钟（Minute）、纬度（Start_Lat）、温度（Temperature）、经度（Start_Lng）、气压（Pressure）、湿度（Humidity）和能见度（Visibility），以及自己添加的道路等级（street_flag）和分簇编号（cluster_id）。

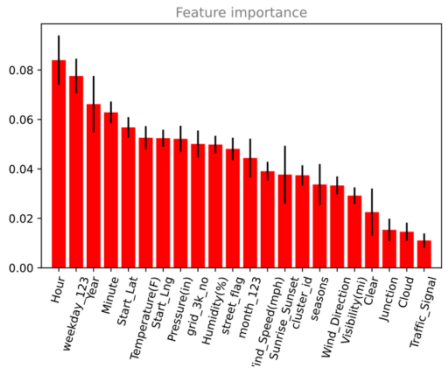


图 3 基于 ERT 的特征重要程度计算

Fig.3 Calculation of feature significance based on ERT

1.3.2Pearson 相关性分析

对特征变量与目标变量之间的相关性强度进行了分析。从特征重要性列表中选取 Top-23 特征变量集。计算与目标变量的皮尔逊积矩相关系数（Pearson's r），即与的协方差与标准差的乘积之比，计算公式如下：

$$r = \rho_{X_i,Y} = \frac{E(X_iY)-E(X_i)E(Y)}{\sqrt{E(X_i^2)-E(X_i)^2}\sqrt{E(Y^2)-E(Y)^2}} \quad (13)$$

图 4 展示了 Pearson 相关性计算结果，总体而言，目标变量 Severity 与给定中的特征变量相关性都不高，其中与 Year、Month 和 Hour 等时序特征变量呈较大的负相关，Pressure 和 Humidity 与目标变量呈较大正相关关系，而图中未涉及的其他特征变量，与特征变量并无相关性关系。

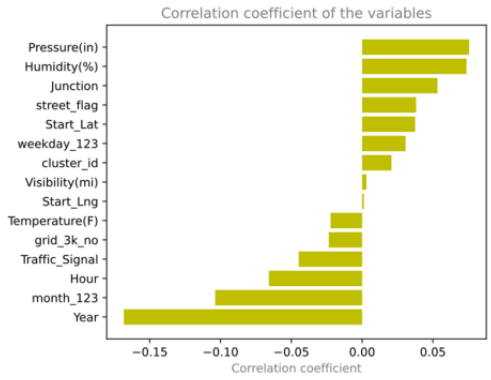


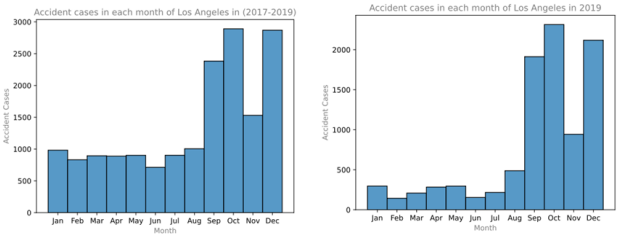
图 4 特征变量与目标变量的相关性

Fig.4 Correlation of the feature variable with the target variable

1.3.3 基于统计学的关联性分析

（一）时间数据分析

对 2017-2019 年月事故量进行统计分析，发现 10 月份最容易发生交通事故，其次是 12 月份和 9 月份，而每年的 2 月跟 7 月事故发生量最少，如图 5 所示。



（a）2017-2019 年 （b）2019 年

(a) 2017-2019 (b) 2019

图 5 月交通事故量直方图

Fig.5 Histogram of monthly traffic accidents

对数据总体按周内和周末统计汇总，如图 6 所示，最左折线图显示了一周内的平均事故严重程度变化，发现周末平均事故严重程度比周内要高；另两幅图分别展示了一周七天的事故量折线以及柱状统计图，事故在周内比周末多出约 87%。

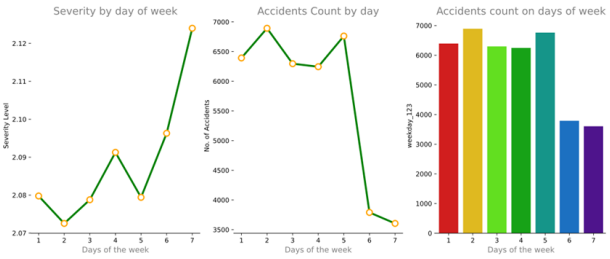
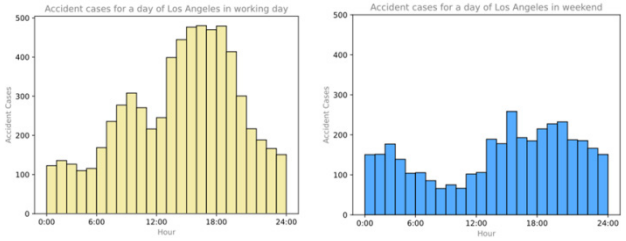


图 6 周内和周末交通事故量

Fig.6 Traffic accidents in the week and on weekends

对周内和周末不同时段发生的事故量进行统计，如图 7 所示，在周内，一天中发生的事故在 14:00~20:00 时间范围内达到最大峰值。而周末一天的事故分布更分散，没有形成很明显的峰值。



（a）周内 （b）周末

(a)weekday (b)weekend

图 7 不同时段事故量统计

Fig.7 Statistics of accidents in different time periods

根据以上时间关联性分析结果，月份（Month）、星期（Week）、小时（Hour）等特征被选定为交通事故的潜在时域预测特征。

（二）空间数据分析

数据显示，洛杉矶共有 1095 条道路，最多发生 2387 起，而大多数发生事故次数小于 5，34.15% 的事故发生在总量占 1% 的 Top10 道路，图 8 可视化了 2016-2020 五年内交通事故量发生量 TOP10 的道路。

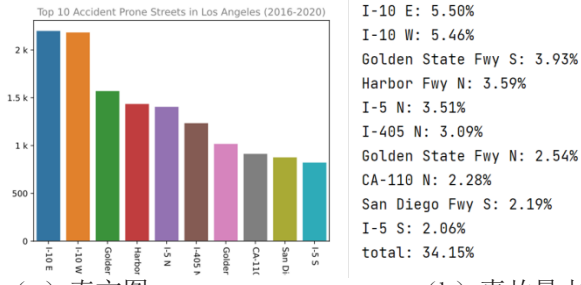
将道路危险等级划分为表 4，划分标准为事故在

道路上的发生频率，其中危险等级为一级的道路，事故发生频率最高，危险程度最高，然后将危险等级构造为一个新的属性“street_flag”。

表 4 道路危险等级划分

Table 4 Classification of road hazard grades

危险等级	一级	二级	三级	四级	五级	六级	七级	八级	九级	十级
事故数量	≥ 2k	≥ 1k	≥ 700	≥ 500	≥ 300	≥ 100	≥ 39	≥ 10	≥ 5	≤ 5



（a）直方图 （b）事故量占比

(a) Histogram (b) Proportion of accidents

图 8 事故发生量 TOP10 道路

Fig.8 TOP10 road of number of accidents

图 9 为交通事故地点缺少 POI 标注的情况，可以发现，绝大多数事故地点都缺少 POI 标注，比如缺少减速带（Bump）POI 高达 99.99%。

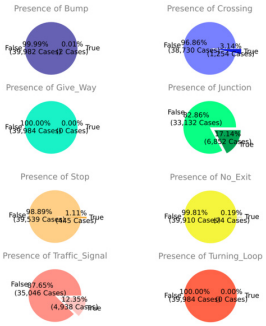


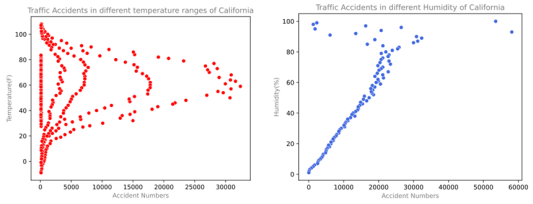
图 9 POI 缺少情况

Fig.9 POI missing situation

根据以上空间关联性分析结果，将道路危险等级（street_flag）、有无交通信号灯（Traffic_Signal）、有无分叉口（Junction）等 POI 标注和 3.2.5 构造的集群编号（cluster_id）选定为特征输入，作为交通事故的潜在空间预测特征。

（三）环境数据分析

为更好的研究温湿度与事故量之间的相关性关系，对整个加州的数据进行了散点表示，如图 10 所示。从（a）可以看出，事故在 50F ~70F 温度范围内发生最多，低于 20F 和高于 9 F，发生量很少。从（b）可以看出湿度与事故数量之间存在一定的正关系，随着湿度升高，交通事故的发生量也随之升高，但是当湿度增至 80% 以上时，这种相关性便消失了。



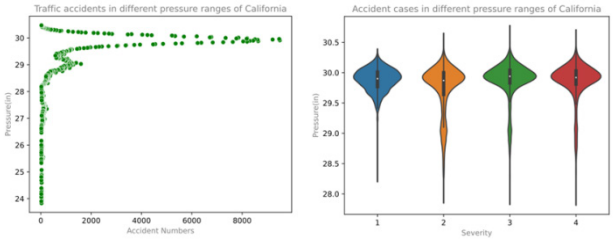
（a）温度 （b）湿度

(a) Temperature (b) Humidity

图 10 温湿度与事故量的相互关系

Fig.10 The interrelationship between temperature and humidity and accident volume

通过散点图矩阵查看气压与交通事故之间的相关关系，如图 11 所示，从图（a）中可以看出，气压大小在 30in 左右最易发生交通事故。图（b）以小提琴图的形式展示了不同严重程度的交通事故在气压下的整体分布，可以看出，所有不同程度的交通事故都集中分布在 29.5in~30.3in，Severity-2 在 29in 处有一个较为明显的峰值，其中位数最低，四分位范围最大。



（a）气压散点图 （b）气压分布

(a) Air pressure scatter diagram. (b) Air pressure distribution

图 11 气压与事故量的相互关系

Fig.11 The interrelationship between air pressure and accident volume

根据以上环境关联性分析结果，将温度（Temperature）、湿度（Humidity）、气压（Pressure）等环境因素选定为模型特征输入，作为交通事故的潜在环境预测因子。

1.3.4 特征选择

根据上述数据分析结果，从空间、时间、POI 和环境四个维度，在改后数据集中选取 26 个特征用于交通风险预测，如表 5 所示。

表 5 特征选取

Table 5 Feature selection

类型	特征			
空间	Start_Lat	Start_Lng	cluster_id	street_flag
时间	Year	month_123	weekday_123	Hour
	Minute	Sunrise_Sunset		

POI	Crossing	Junction	Traffic_Signal	Station
	Stop			
环境	Temperature(F)	Humidity(%)	Pressure(in)	Visibility(mi)
	Wind_Direction	Wind_Speed(mph)	Cloud	Clear
	Rain			

2 交通风险预测模型

2.1 算法总体架构

交通风险预测模型算法架构如图 12 所示，模型由“交通事故二分类器”、“事故严重程度多分类器”以及“交通风险大小评估函数”三部分组成。

其中，交通事故二分类器用于预测某路段在某种情境下是否发生交通事故，在训练阶段，输入为从阳性样本集和 Unlabeled 样本集中提取的特征变量 y，输出为交通事故为 TRUE 的概率。

事故严重程度多分类器用于判断预测某地发生交通事故的严重程度，输入为从阳性样本集中获取的特征变量 x，输出为代表事故严重程度的数字 2、3、4。

交通风险评估函数根据事故发生概率和事故严重程度对某区块的风险大小进行评估，输出为风险数值。

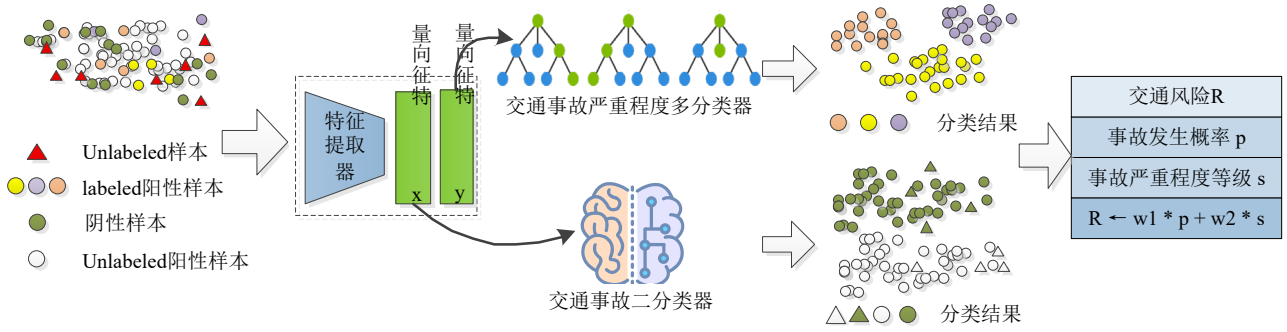


图 12 算法架构图

Fig. 12 Algorithm architecture diagram

2.2 交通事故二分类器

交通事故二分类器设计流程如图 13 所示。首先，进行数据分析并构造特征变量；然后，采用特征重组的方式，构造 Unlabeled 样本集，并经过 PU 学习，实现对阴性样本集的可信划分，在此基础上进一步构造出训练集和测试集；最后构建二分类模型，并通过 k 折交叉验证与网络搜索实现对最优模型参数的选择。

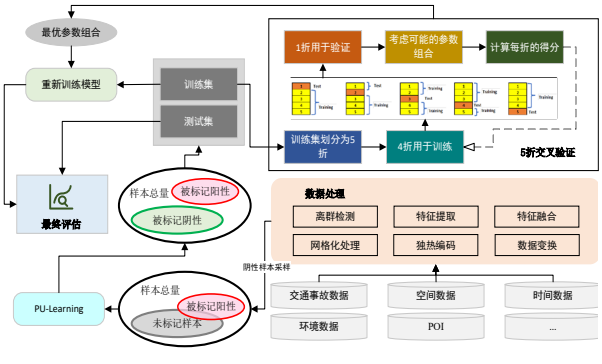


图 13 交通事故二分类器设计图

Fig. 13 Design drawing of traffic accident second classifier

（一）Unlabeled 样本集构建

当前数据集中所有的样本都为阳性数据，若要对某地交通事故发生的概率进行判断的话，总体中必须得包含负样本，所以要训练一个二分类模型，首先需要对负样本进行采样，本文采用特征重组的负样本采样方式，即随机选择数据集中的不同字段的值来生成一条记录，只要该记录未出现在数据集跟阴性列表中，则将其添加到阴性列表，样本构建过程如下：

1）数据纵向切分

首先，根据时空敏感性，将数据集纵向分割为时间敏感性数据帧（Time sensitive data-frame，TSD）和空间敏感性数据帧（Spatial sensitive data-frame，SSD），TSD 包含了时间、以及该时间条件下对应的温度、湿度等环境变量，其他部分则被划分进了 SSD。一方面可以将异质性的强关联的数据进行拆分，另一方面可以很好地保留样本的时空特性。

2）随机抽样

然后，打乱这两个数据帧，通过随机抽样的方式，分别随机抽取两个数据帧的数据子集，并进行数据组合，得到一个新的样本。

3）可靠性检测

最后，检查该样本是否被包含在阳性样本里，如果没有被包含，则将其添加到未标记（Unlabeled）样本集，否则丢弃，各样本集之间的关系维恩图如图 14 所示。

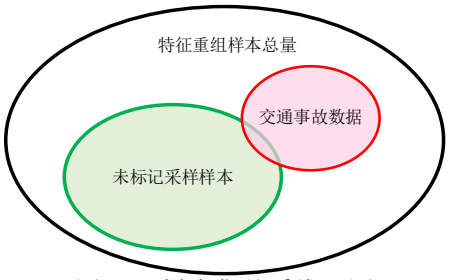


图 14 样本集关系维恩图

Fig. 14 A Venn diagram of sample-set relationships

2.3 事故严重程度多分类器

首先，对数据集进行过采样和欠采样处理，解决数据集本身不平衡性问题。然后，结合交叉验证与网络搜索方法，比较了逻辑回归、KNN、决策树和随机森林等机器学习算法的预测性能，决定选用随机森林模型来进行实现。

图 15 为样本集交通事故严重程度占比分布情况，可以看到，Severity-2 的样本数量占比高达 93.2%，远远高于其他两类别的样本数量，数据的类别非常不平衡。如果在这种数据分布下创建交通严重程度分类模型，会产生虚假的高分类性能。

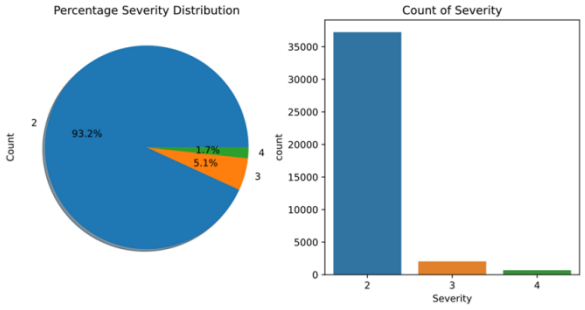


图 15 原数据集分布

Fig. 15 Distribution of the original dataset

将 Severity-2 的样本数量随机欠采样至 20000，即从数据集中随机选择 20000 条 Severity 等于 2 的样本；将 Severity-3 和 Severity-4 的数目过采样至 20000，即分别从数据集中随机选取 Severity 等于 3 和 4 的样本，并对其进行复制来扩展样本容量，直至数量达到 20000，采样过后的类别占比情况如图 16 所示。

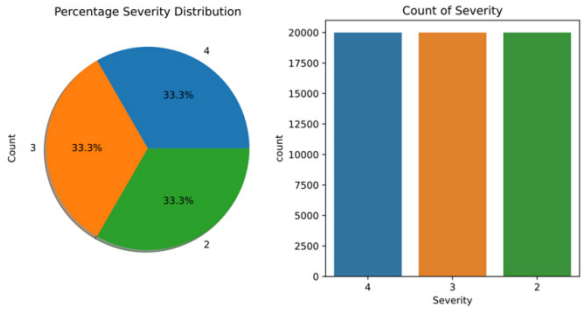


图 16 采样过后数据集分布

Fig. 16 Distribution of the data sets after sampling

2.4 交通风险评估函数

结合事故发生的概率和事故严重程度，构造如下风险评估函数，各参数含义如表 6 所示：

$$R(t) = w1 * p(f(\cdot), t) + w2 * s(g(\cdot), t) \quad (14) \quad (14)$$

表 6 参数含义

Table 6 Parameter meaning

参数	含义
	在时间下发生风险的大小
	用于预测事故发生概率的特征参数
	交通事故发生概率
	用于预测事故严重程度的特征参数
	事故严重程度等级
	权重参数，调节概率在风险值中的影响
	权重参数，调节严重程度在风险值中的影响

2.5 道路离散化

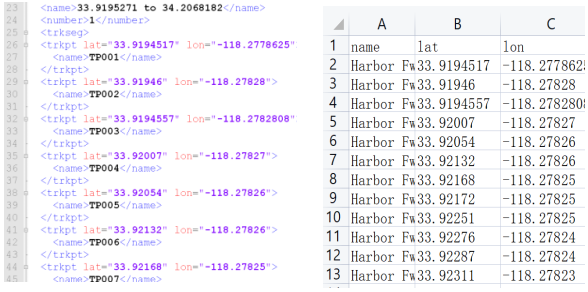
利用上述交通风险预测模型，可以对城市道路交通事故进行实时的风险预测，区域中的每条道路都可以被离散为固定长度的小路段，在每个路段中心处，模型可以结合给定的时空、环境等信息对是否发生交通事故进行预测，并完成风险值的输出。

（1）道路信息抓取

首先，通过 Google Roads API，对数据集中涉及的 1095 条道路进行道路信息抓取。

（2）坐标点解析

将抓取的 Google 道路信息换为 GPX 文件，进一步将其解析到 csv 文件，如图 17（b）所示，其中“name”列为道路名称，“lat”和“lon”分别为路径点的维度和经度。为减少模型计算量的同时，尽可能的保证道路全覆盖，采用均匀采样的方式提取该解析数据的 1/3。



（a）GPX 提取 （b）GPX 解析

图 17 坐标点解析

Fig. 17 Coordinate point resolution

3. 实验分析

3.1 事故严重程度多分类器性能分析

经数据平衡处理后，分别使用逻辑回归、KNN、决策树和随机森林等机器学习模型对交通事故严重程度进行预测，训练集随机选取 75%，剩余 25% 用于测试，性能评价指标选择准确率（Accuracy）、查准率（Precision）、查全率（Recall）、F1 分数（F1-Score），表 7 列出了各模型在该评价指标上性能情况，与基准

模型相比，随机森林模型具有最优的性能。

表 7 分类器性能指标对比

Table 7 Comparison of classifier performance indicators

模型	准确率	查准率	查全率	F1-Score
逻辑回归	0.520	0.520	0.520	0.520
KNN	0.874	0.890	0.870	0.870
决策树	0.915	0.920	0.920	0.910
随机森林	0.971	0.970	0.970	0.970

图 18 和 19 分别为各模型的 ROC 曲线和 PR 曲线，从中可以看出，随机森林的 AUC 值为 0.998，高于逻辑回归（0.706）、KNN（0.968）、决策树（0.977）。这说明，在 AUC 评价指标上，随机森林模型的分类效果最好。

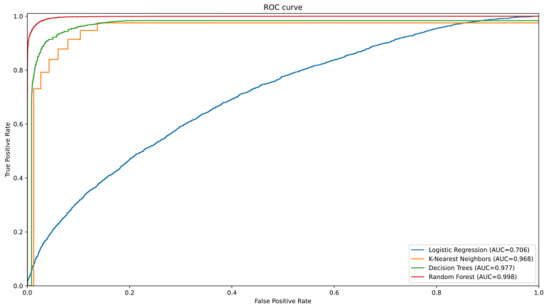


图 18 事故严重程度分类器 ROC 曲线

Fig.18 Accident severity classifier ROC curve

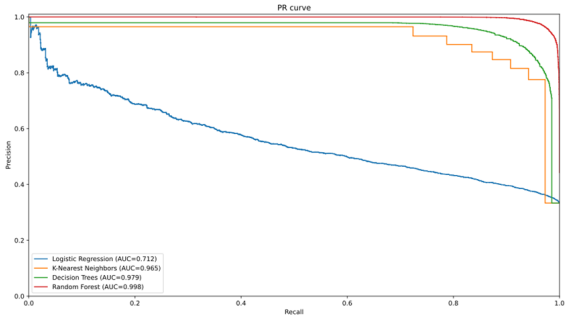


图 19 事故严重程度分类器 PR 曲线

Fig.19 The PR curve of the accident severity classifier

与现有预测模型相比，Castro 等人[47]采用贝叶斯网络来进行事故严重程度预测，模型准确率为 0.8159，查准率为 0.7239，查全率为 0.7239，F1 分数为 0.723。文献[48]将事故严重程度分为有受伤人员和无受伤人员，并基于 LightGBM 模型建立了事故严重程度二分类模型，模型准确率、AUC 和 F1 分数分别为 0.673、0.698 和 0.790；文献[49]建立了道路事故严重程度三分类 RF 模型，其准确率、F1 分数、AUC 值分别为 0.853、0.691 和 0.80；文献[50]通过建立人工神经网络分类模型，准确率和 AUC 分别为 0.746 和 0.752；与本文采用同样数据集的文献[51]，结合随机森林和卷积神经网络提出了一个集成模型 RFCNN，在将所有特征变量作为输入时，模型准确率、查准率、查全率和 F1 分数分别为 0.812、0.842、0.864 和 0.853，而提取 20 个显著的特征输入后，模型具有 0.991 的准确率、0.974 的查准率、0.986 的查全率和 0.980 的 F1 分数，性能略高于本模型，但是本模型以较低的计算复杂度实现了非常高的模型表示能力，总结对比如表 8 所示。

表 8 分类器性能指标对比

Table 8 Comparison of classifier performance indicators

模型	准确率	查准率	查全率	F 1 - Score	AUC
Bayes[47]	0.816	0.724	0.724	0.723	-
LightGBM[48]	0.673	-	-	0.790	0.698
RF[49]	0.853	-	-	0.691	0.800
神经网络 [50]	0.746	-	-	-	0.752
RFCNN[51]	0.991	0.974	0.986	0.980	-
本模型	0.971	0.970	0.970	0.970	0.998

采用消融实验，对经数据处理和数据分析所构造的“cluster_id”、“weekday_123”、“grid_3k_no”及“street_flag”等四个属性的有效性进行验证，评估它们对模型性能的影响，构造特征列表；然后从特征列表中分别删除的不同子集，生成新的特征列表；使用训练新的随机森林分类模型，并查看其性能报告。

通过上述验证过程，得到模型对比评估结果如表 9 所示，对分类报告结果保留小数点后三位小数。

表 9 模型对比评估结果

Table 9 Model comparison and evaluation results

特征列表	准确率	查准率	查全率	F1-Score
	0.972	0.973	0.972	0.972
	0.964	0.966	0.964	0.964
	0.955	0.959	0.955	0.955
	0.966	0.968	0.966	0.966
	0.967	0.970	0.967	0.967
	0.965	0.967	0.965	0.9654
	0.966	0.967	0.966	0.966
	0.963	0.966	0.963	0.963
	0.954	0.958	0.954	0.954
	0.963	0.965	0.963	0.963
	0.956	0.959	0.956	0.956
	0.950	0.953	0.950	0.949

从上述结果可以看出，所构造的四个属性对于提高模型的分类性能都产生了积极的贡献，如果舍弃其

中任何一个或几个属性，都会导致模型性能的下降。最差情况下，模型的准确率由 97.2% 下降至 95%，查准率由 97.3% 下降至 95.3%，查全率 97.2% 下降至 95%，F1 分数由 97.2% 下降至 94.9%。通过此消融实验，证实了所构造属性的有效性。

3.2 交通事故分类模型性能分析

模型性能报告如图 20 所示，“support”表示当前分类在测试集中的样本数量，即 class 0 类别在测试集中总量为 10003，class 1 为 9808，二者数量比约为 1:1，模型的准确率、查准率、查全率和 F1 分数分别为 0.75、0.74、0.77 和 0.73。

[PU Bagging (Random Forest) algorithm] classification_report:

	precision	recall	f1-score	support
0	0.72	0.78	0.74	10003
1	0.76	0.76	0.72	9808
accuracy			0.75	19811
macro avg	0.74	0.77	0.73	19811
weighted avg	0.74	0.77	0.73	19811

图 20 交通事故二分类器分类报告

Fig.20 Traffic accident taxfier classification report

模型的 ROC 曲线如图 21 所示，所提的基于 PU Bagging 算法的随机森林模型 AUC 值为 0.878，与逻辑回归、KNN 等 8 个基准模型相比，具有最高 AUC 值，在数据不平衡问题的处理上，可靠性阴性样本的选择至关重要。

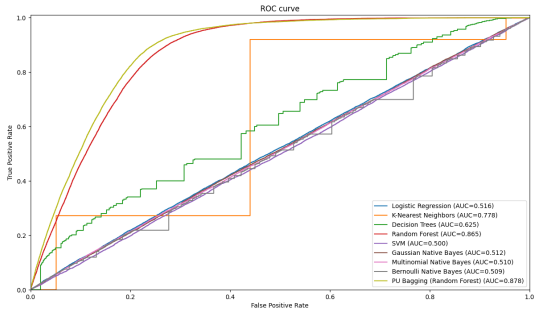


图 21 交通事故二分类器 ROC 曲线

Fig.21 The ROC curve of the traffic accident secondary classifier

与现有模型相比，文献[52]基于卷积神经网络提出了一种用于实时交通事故预测的二分类模型，预测精度为 0.785；文献[53]将交通事故预测问题构造为回归问题，基于长短时记忆网络的深度学习方法，对特定路段风险进行实时预测，该模型综合考虑了气象信息、POI、时间和车流量等影响因素，共构建了 76 个特征，模型精准度、查全率和 F1 分数分别为 0.723、0.773 和 0.736；文献[54]构建了一个基于递归神经网络的深度学习模型，来对交通事故风险进行预测，平均绝对误差和均方根误差分别为 0.014 和 0.034，但是该模型对影响事故的数据分析不够，而且预测粒度为每隔 3 天，实时性较差。

4 结束语

基于历史真实交通事故数据，对道路交通事故与动态影响因素之间的内在联系进行深入挖掘，对诱发交通事故发生的影响因素进行了全面考虑，提升了模型的预测能力。进一步设计了结构化交通风险实时性预测模型，将交通风险分解为事故发生概率与事故严重程度等级两个层次，更加细致地捕捉到各个因素之间的相互作用，使得这些复杂关系能够被更好地建模和识别，对交通事故风险的预测提供可靠依据。未来的研究可以在此基础上进一步优化模型，探索更多影响交通安全的因素，以持续提升交通风险管理的科学性与有效性。

参考文献

[1]Beg A, Qureshi A R, Sheltami T, et al. UAV-enabled intelligent traffic policing and emergency response handling system for the smart city[J]. Personal and Ubiquitous Computing, 2021, 25: 33-50.

[2]Zhang J, Jiahao X. Cooperative task assignment of multi-UAV system[J]. Chinese Journal of Aeronautics, 2020, 33(11): 2825-2827.

[3] 张祥银，夏爽，张天．基于自适应遗传学习粒子群算法的多无人机协同任务分配[J/OL]. 控制与决策:1-9[2022-10-08].DOI:10.13195/j.kzyjc.2022.0240.

[4]Shima T, Rasmussen S J, Sparks A G, et al. Multiple task assignments for cooperating uninhabited aerial vehicles using genetic algorithms[J]. Computers & Operations Research, 2006, 33(11): 3252-3269.

[5]Zhu M, Du X, Zhang X, et al. Multi-UAV rapid-assessment task-assignment problem in a post-earthquake scenario[J]. IEEE access, 2019, 7: 74542-74557.

[6]Xiao K, Lu J, Nie Y, et al. A benchmark for multi-UAV task assignment of an extended team orienteering problem[J]. arXiv preprint arXiv:2009.00363, 2020.

[7]Singh A, Baghel A S. A new grouping genetic algorithm approach to the multiple traveling salesperson problem[J]. Soft Computing, 2009, 13: 95-101.

[8]Srinivas M, Patnaik L M. Genetic algorithms: A survey[J]. computer, 1994, 27(6): 17-26.

[9] 陈长征，王楠．遗传算法中交叉和变异概率选择的自适应方法及作用机理[J]. 控制理论与应用,2002(01):41-43

作者简介：陈光晓（1996-）男，汉，山东潍坊，硕士研究生，助理工程师，研究方向：信息系统开发。