引文格式：吴国超,鲍文月,李萌. 一种交通事件向量化表征与分类方法[J]. 地理空间信息, 2026, 24(2):118-120. doi:10. 3969/j.issn. 1672-4623.2026.02.025

一种交通事件向量化表征与分类方法

吴国超[1]，鲍文月[2]，李　萌[1]
（1. 常州市武进规划勘测设计院，江苏 常州 213000；2. 常州市新北自然资源和规划技术保障中心，江苏 常州 213000）

摘　要：随着城市交通系统的复杂性日益增加，交通安全问题受到广泛关注，对交通事件进行快速自动化分析和处理需求日益迫切。通过将交通事件的文本描述转换为向量形式，采用Word2Vec的Skip-gram模型来生成高质量的词向量，并基于这些向量构建支持向量机（SVM）分类模型，对事件的严重性和类型进行自动分类。通过对常州市武进区湖塘镇2021—2023年交通事故数据分析，实现了对该地区3 a交通事故的分类与分析，相较传统的交通事件统计分析方法，本方法语义捕捉能力更强，分类效率更高，对于交通事件的全面分析，优化事件响应和预防措施具有显著优势。

关键词：交通事件；向量化表征；事件分类
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2026）02-0118-03

Vector Representation and Classification Method of Traffic Events

WU Guochao[1], BAO Wenyue[2], LI Meng[1]
(1. Changzhou Wujin Planning and Surveying Design Institute, Changzhou 213000, China;
2. Changzhou Xinbei Natural Resources and Planning Technical Support Center, Changzhou 213000, China)

Abstract: With the increasing complexity of urban traffic systems, traffic safety issues have gained widespread attention, and there is an urgent need for rapid and automated analysis and processing of traffic events. Converting the textual descriptions of traffic events into vector forms, we used Skip-gram model of Word2Vec to generate high-quality word vectors, and based on these vectors, constructed a support vector machine (SVM) classification model to automatically categorize the severity and type of events. By analyzing traffic accident data from 2021 to 2023 in Hutang Town, Wujin District, Changzhou City, Zhejiang Province, China, we classified and analyzed traffic accidents in the region over three years. Compared to traditional statistical analysis methods of traffic events, the proposed method demonstrates stronger semantic capture capabilities and higher classification efficiency, providing significant advantages for comprehensive analysis of traffic events and optimizing response and prevention measures.

Key words: traffic event, vector representation, event classification

随着城市化进程的加快，交通系统的复杂性日益增加，交通安全问题也随之成为社会关注的焦点。交通事故不仅对个人和家庭造成严重影响，还会对社会经济和城市运行效率带来负面影响。因此，如何有效利用现有的交通事故数据来预防和减少事故发生，是当前交通管理和城市规划领域面临的重要问题[1-6]。

本研究探索了Word2Vec技术在交通事故向量化表征与分析中的应用。通过将交通事故的文本描述转换为向量形式，采用Word2Vec的Skip-gram模型来生成高质量的词向量，并基于这些向量构建SVM分类模型，对事故的严重性和类型进行自动分类。

1　研究区域概况

研究区域是位于武进区中心城区的湖塘镇，该区域外围工厂分布密集，镇区内业态功能复杂，是武进区的生活服务核心，中心区域以居住用地为主，用地功能丰富，成熟度较高，区域外围为产业园区，工厂密集，居住区与周边工业区域联系紧密[7-8]。

交通方面，该区域以通勤为主体，具有一定潮汐特性，早高峰流向以西向东、北往南为主，往工业区方向，晚高峰以东往西、南往北为主。区域内支路交通流量较低，高峰期交通量集中于常武路、夏城路、花园街、人民路、广电路、长虹路等主次干路。

2　研究方法

2.1　总体技术路线

本文提出的一种交通事件向量化表征与分类方法以Word2Vec算法为核心，主要包括数据准备与预处

项目来源：自然资源部国土卫星遥感应用重点实验室开放基金资助项目（KLSMNR-G202301）。
第一作者简介：吴国超（1990—），硕士，高级工程师，从事自然资源数据处理与地理信息系统研发工作，E-mail：wuguochaowgc@126.com。

理、构建事件序列、模型训练与向量化、事件聚类分析4个主要环节，技术路线如图1所示。

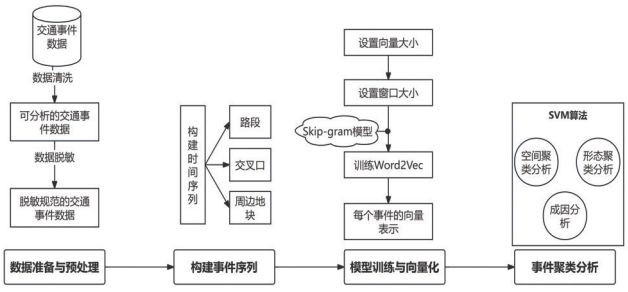


图1　交通事件向量化表征与分类方法技术路线

2.2　Word2Vec方法与Skip-gram模型

Word2Vec是一种高效的词嵌入模型，通过深度学习技术将词汇映射到一个连续的向量空间中，以捕捉词与词之间的语义关系。该模型由Google于2013年提出，主要包括2种架构：Skip-gram模型和Continuous Bag of Words（CBOW）模型[9]，本文采用Skip-gram模型，因其在处理稀疏数据和大规模语料时具有显著优势[10]。

Skip-gram模型的目标是通过给定一个中心词来预测其上下文词。假设有一个词汇表V，给定一个中心词，其目标是最大化该词与其上下文词（k为窗口大小）出现的概率[11]。模型试图最大化式（1）目标函数：

$$J=\sum_{t=1}^{T}\sum_{-c\leqslant k\leqslant c}\log P\left(w_{t+k}/w_t\right) \tag{1}$$

式中，T为语料库中的总词数；c为上下文窗口大小。为计算上下文词的条件概率$P\left(w_{t+k}|w_t\right)$，使用softmax函数进行规范化，函数如式（2）所示：

$$P\left(w_j|w_i\right)=\frac{e^{\boldsymbol{v}_{w_j}\cdot\boldsymbol{v}_{w_i}}}{\sum_{w=1}^{|V|}e^{\boldsymbol{v}_{w}\cdot\boldsymbol{v}_{w_i}}} \tag{2}$$

式中，$\boldsymbol{v}_{w_i}$和$\boldsymbol{v}_{w_j}$分别为词w_i和w_j的向量，通过优化该目标函数，模型学习到每个词的向量表示。

通过训练后，词向量的维度通常设置为100~300维[12]，使得语义相近的词在向量空间中相互靠近。

2.3　支持向量机（SVM）分类模型

支持向量机（SVM）是一种广泛应用于分类与回归分析的监督学习模型。其基本思想是通过构造一个最优超平面，将不同类别的数据点分开，以实现对未知样本的分类[13]。SVM可以进行小样本学习，且具有较好的鲁棒性[13]。

SVM的目标是找到一个超平面H来分隔不同类别的数据点。假设有两类样本$\left\{\left(x_i,y_i\right)\right\}_{i=1}^{n}$，其中$x_i\in R$表示特征向量，$y_i\in\{-1,1\}$表示类别标签。超平面可以用式（3）方程表示[14]：

$$H:w^{\mathrm{T}}x+b=0 \tag{3}$$

式中，$\boldsymbol{w}$为超平面的法向量；b为偏置项。SVM的目标是最大化分类间隔，即从超平面到最近的数据点的距离。该距离可以式（4）表示：

$$\mathrm{Margin}=\frac{2}{\|w\|} \tag{4}$$

要找到具有最大间隔的划分超平面，也就是要找到满足式（5）、式（6）中约束的参数w和b：

$$\text{minimize}\ \frac{1}{2}\|w\|^2 \tag{5}$$

$$\text{subject to}\ y_i\left(w^{\mathrm{T}}x_i+b\right)\geqslant 1,\forall i \tag{6}$$

在本方法中，基于Word2Vec生成的交通事件向量，采用SVM对事故的严重性和类型进行分类。通过对模型进行训练和验证，期望提高交通事故分析的准确性和效率[14]。

3　交通事件向量化表征与分析实例

针对常州市武进区湖塘镇近3 a（2021—2023年）交通事故信息简报（共计1 778条），基于Word2Vec的向量化表征与分类方法，对湖塘镇交通事故进行了向量化与事件聚类分析，验证方法的有效性。

3.1　向量化表征与分类

数据清洗：对原始数据进行清洗，去除重复记录、无效信息和缺失值，以确保数据的完整性和可靠性，清洗后得到1 536条有效交通事故信息。

文本预处理：删除非文本字符，如标点符号和数字，以保持文本的纯净性。使用中文分词工具对事故描述进行分词，确保后续分析的有效性。剔除常见的停用词，提升重要词汇在向量化过程中的权重。

构建事件序列：将清洗后的数据按时间顺序整理，形成交通事件的时间序列，便于后续分析和模型训练。从事件描述中提取关键信息，如事故类型、天气情况、路况等，以便在向量化过程中增强模型的表现力。

Word2Vec模型训练：采用Word2Vec的Skip-gram模型对处理后的事故描述进行训练，得到词向量（100维），每个维度的值表示该词在某个特定方向上的“语义分布”，部分词向量如表1所示。

Word2Vec模型会将语义上相似的词（例如“小

表1 部分词向量值

词名	向量值
小客车	[0.032 6,−0.125 2,0.084 2,...,−0.215 7,0.102 3,0.878 9]
中型客车	[0.040 2,−0.124 7,0.088 6,...,−0.232 3,0.102 1,0.876 5]
电动自行车	[0.048 8,0.075 6,−0.24 5,...,0.734 6,0.534 7,0.116 5]
自行车	[0.050 7,0.069 8,−0.23 1,...,0.755 4,0.561 9,0.117 7]

客车”与“中型客车”）映射到相近的向量空间，通过计算词向量之间的余弦相似度即可评估词与词之间的关系。

文档向量生成：使用TF-IDF对词向量进行加权，将每个词向量与它的TF-IDF权重相乘，得到加权的词向量，再计算文档中所有词的加权词向量的平均值，生成每个事故描述的文档向量。这一过程将关键字的权重纳入考量，能够突出与交通事故更相关的关键词，提高交通事故文档向量表征的质量。

3.2 交通事故分类

在对研究区域近3 a的交通事故数据进行向量化表征后，下一步是对这些事故进行分类，以揭示事故的类型和严重性。具体步骤如下：

1）分类任务定义：事件分类主要分为三类，即严重性分类；形态分类；成因分类。

2）数据准备与特征选择：使用之前生成的文档向量作为特征输入，确保每个事故的描述信息都被充分利用。同时，将事故的具体类别进行标签编码，形成一个可供模型学习的训练数据集。

3）SVM模型训练：将完整数据集随机划分为训练集（70%）和测试集（30%），以确保模型的训练与评估独立进行。

4）模型训练：在训练集上使用SVM模型进行训练，优化模型参数以提高分类精度。

5）模型分类效果评估：在测试集上进行模型评估，得到模型分类准确率（分类正确的样本占总样本的比例）、召回率（正确分类的正样本占实际正样本的比例）与*F*1分数（调和平均数），模型评估效果如表2所示。

表2 模型分类效果评估结果

交通事件分类模型	准确率	召回率	*F*1分数
严重性分类模型	0.917	0.926	0.921
形态分类模型	0.875	0.894	0.884
成因分类模型	0.843	0.867	0.855

3.3 交通事故分析

基于分类结果，进一步分析研究区域近3 a交通事故情况。分别对一般程序事故与亡人事故两类事故进行空间聚类分析，得到交通事故空间聚类分析图如图2所示。

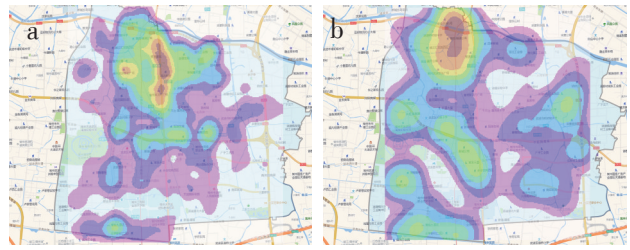


图2 研究区域2021—2023年交通事故空间聚类热力图

审图号：苏D(2019)1

空间聚类分析结果表明研究区域近3 a一般程序事故主要集中于常武路沿线的城镇区域，湖塘纺织城周边事故量相对较大（图2a）。亡人事故共发生主要集中于湖塘镇区西北角，其中花园街大桥、东部工业区、纺织城等片区事故发生率较高（图2b）。

对事故的形态、事故成因、死亡事故成因也进行了聚类分析，发现事故主要形态为中小客车与非机动车、中小客车与中小客车、非机动车与非机动车，三者占比接近70%。事故成因排名前列有多项与非机动车相关，包括违反交通信号灯、逆行、过街未下车推行等不规范交通行为，占比超50%。违反交通信号灯为主要的亡人事故成因之一，行人、非机动车及机动车违反信号灯行为造成的亡人事故达到68起，超过30%。

值得注意的是，事故特征与研究区域出行特征及车辆自身比例构成有关，根据《常州市武进区综合交通调查报告（2024年）》，常州市武进区慢行出行方式（步行、自行车、电动自行车）占比69.2%，其中电动自行车占比49.0%。非机动车比例较大与该地区的慢行出行方式的增加存在一定的关联，后续的交通事故分析及措施建议需要考虑这些因素。

3.4 方法效果分析

与传统方法相比，本方法优势主要体现在以下3点：①自动化特征提取。②语义理解与上下文感知。③向量化方法的准确性较传统方法更高。

4 结 语

本文提出了一种交通事件向量化表征与分类方法，基于Word2Vec算法对常州市武进区湖塘镇近3 a的交通事故信息进行了向量化表征与分析，并利用SVM模型对交通事故进行分类，研究实现了对事故严重性和类型的准确识别。评估结果显示，模型在准确

（下转第124页）