DOI:10.16749/j.cnki.jecjtu.2014.05.011



文章编号:1005-0523(2014)05-0050-06

# 基于贝叶斯网络的交通事件持续时间预测

郑长江[1],葛升阳[1],郑树康[2]

(河海大学 1.土木与交通学院,江苏 南京 210098;2.物联网学院,江苏 常州 213022)

**摘要**:随着数据采集手段的不断提高和相关研究技术的发展,基于数据挖掘的模型逐渐成为交通事件持续时间研究的主要方向。根据荷兰交通部门提供的交通事件采集数据,进行分类和预处理,观察事件持续时间的频数图,并根据相关的研究按照事件典型的类别把采集的数据进行分类。使用主成分分析和逐步回归提取出显著性的影响因子,利用数据挖掘软件WEKA建立贝叶斯网络模型,用数据集中80%的数据进行学习建模,20%的数据作为测试集来检测模型的预测效果,并做出性能评价。实验结果表明,与同类数据集的其他预测方法相比,贝叶斯网络模型对于变数众多,随机性特别大的交通事件,预测精度较高,证明贝叶斯网络模型的算法是具有一定优越性和实用价值。

**关键词**:城市交通;交通事件持续时间;贝叶斯网络模型;数据集分类;影响因子提取;WEKA

**中图分类号**:U491.1 **文献标志码**:A

交通拥堵会造成经济损失、能源过量消耗、出行时间延误,环境污染等可估计和不可估计的损失,弊端之大,不言而喻。据不完全统计,在美国,60%的城市道路交通拥堵是由偶发性的交通事故、车辆抛锚和货物掉落等引起的[1]。

不同于每日高峰时段的常发性交通拥堵,无法预知的偶发性交通事件也是导致交通拥堵的重要因素,并且此类拥堵相比常发性拥堵更容易引发二次事故。这里交通事件指的是不可预知的偶发性事件,包括交通事故、碰撞、抛锚,车辆着火,道路施工、天气情况等。美国在2003年的统计显示,全美范围内车辆碰撞发生600万次,导致死亡人数42 000人,受伤人数29 000 000人,经济损失总值约2 306亿美元,相当于美国国民生产总值的2.3%[2]。

国内外很多研究者致力于交通事件持续时间的预测,每一个研究使用的数据集不同,事件变量不同,样本容量也不同。总结这些方法,有以下几类:时间序列模型,线性回归模型,非参数回归模型,基于概率分布的预测方法,基于条件概率的预测方法,决策树预测模型和人工神经网络,Cox Regression模型,多元回归分析,模糊逻辑预测[3-5]等。考虑众多方法的优缺点,将贝叶斯网络预测模型用于城市道路交通事件持续时间的研究,并对事件总数据集进行了分类和影响因子的提取,提高了精度。

## 1 交通事件持续时间的定义

一般,交通事件持续时间包括4个重要的组成部分,并且各部分相互独立,即:事件的发现时间,事件响应时间,事件清除时间和交通恢复时间,具体如图1所示。

事件发现阶段:从交通事件发生到交通管理者通过各种信息渠道得知发生事件的时间阶段。事件响应阶段:交通事件被确认之后,各方面的营救人员和救援车辆到达现场的时间阶段。事件清除阶段:各方面的救援行动如抢救受伤人员,车道封锁,移除事件车辆以及碰撞碎片等结束以后,道路开始恢复通行能

**收稿日期**:2014-07-18

**基金项目**:江苏省自然科学基金项目(BK2011745)

**作者简介**:郑长江(1966—),男,教授,博士生导师,研究方向为交通规划与管理。

**通讯作者**:葛升阳(1989—),男,硕士研究生,研究方向为交通规划与管理。

中国知网 https://www.cnki.net

力的时间阶段。事件恢复阶段：交通事件被彻底清除后，车辆排队消散直至道路恢复原有的正常通行能力的时间阶段。

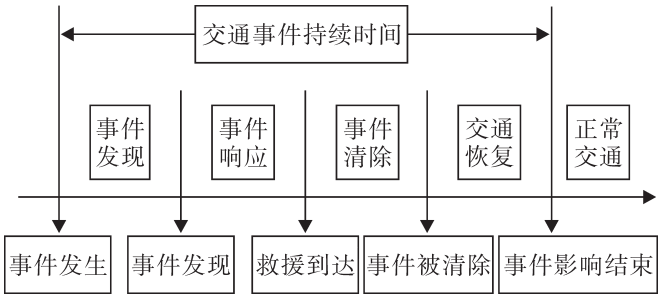


**图1　交通事件持续时间的定义**

**Fig.1　Definition of traffic incident duration**

## 2　贝叶斯网络模型概述

贝叶斯网络的原理，就是统计学上的概率推理，所谓概率推理就是通过一些变量信息来获得其他变量的概率信息。首先，关于贝叶斯网络的相关概率公式介绍如下。

1）条件概率。设$A$，$B$是两个事件，且$P(B)>0$，称$P(A|B)=\dfrac{P(AB)}{P(B)}$为已知事件$B$发生的条件下，事件$A$发生的条件概率。

2）联合概率。若$A$、$B$为两个基本事件，切$P(B)>0$，则有

$$P(AB)=P(B)P(A|B) \tag{1}$$

上式为乘法公式，$P(AB)$称为$A$、$B$的联合概率分布。

3）全概率公式。设$B_1$，$B_2$，$\cdots B_n$是一系列两两互不相容的事件，且有

$$\bigcup_{i=1}^{n}B_i=\Omega,P(B_i)>0,i=1,2,\cdots,n \tag{2}$$

则对任一事件$A$，有

$$P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i) \tag{3}$$

4）贝叶斯公式。根据公式(1)和(2)，可以推导出贝叶斯公式

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(A|B_j)P(B_j)} \tag{4}$$

贝叶斯网络$B=<B_s,B_p>$表示$n$个随机变量$X=\{X_1,\cdots,X_n\}$的联合概率分布，这个网络由2部分组成[6]：

① DAG，即有向无环图，表示网络结构$B_s$。$n$个随机变量在结构图中以节点表示，节点之间的有向边代表了节点之间的相互关系，即变量之间的概率依赖关系。如果$X$节点和$Y$节点之间的有向边是指向$Y$的，那么称$X$为$Y$的父节点，$Y$是$X$的子节点。

② CPT，即条件概率表，用以反映变量之间的相关联系。

$B_p=\left\{P(X_i|\Pi_{X_i}),1\leqslant i\leqslant n\right\}$，其中$X_i$是网络中的节点，$\Pi_{X_i}$是$X_i$的父节点集，若$X_i$没有父节点，则$\Pi_{X_i}=\varnothing$。

根据概率论的原理，贝叶斯网络的联合概率分布

$$P(X_1,\cdots,X_n)=\Pi_{i=1}^{n}P(X_i|X_1,\cdots,X_{i-1}) \tag{5}$$

而贝叶斯网络中，每个节点在其父节点已知的时候是条件独立与其他非子节点的，即

$$P(X_i|X_1,\cdots,X_{i-1})=P(X_i|\Pi_{X_i}) \tag{6}$$

中国知网　https://www.cnki.net

根据公式(5)和(6)可得

$$P(X_1,\cdots X_n)=\Pi_{i=1}^{n}P(X_i|\Pi_{X_i}) \tag{7}$$

## 3 贝叶斯网络模型构建

### 3.1 确定网络节点

贝叶斯网络节点就是刻画所研究对象的一组随机变量集合$X=\{X_1,\cdots,X_n\}$,用$x_i$来表示$X_i$这个随机变量的取值。

贝叶斯网络节点中包含了目标节点和证据节点。目标节点就是我们建立贝叶斯网络模型需要求解的未知变量,证据节点是作为推理的证据输入网络的,其本身可以由观测得到数据。贝叶斯网络是以概率来推理的,所以变量节点的取值就必须是离散型的数据值,最典型的取值形式就是{yes,no}。

### 3.2 确定网络结构

基于评分搜索的网络结构算法是从一个初始网络出发,利用搜素算法修改完善网络结构,再利用评分函数对习得的网络结构打分,然后重复这一步骤,直到找到最优的网络结构为止[7]。那么评分搜索的算法就有2部分构成:① 评分函数;② 搜索算法。

定义了评分函数之后,贝叶斯网络学习的问题就演化为了搜索方法问题。通过拟定的搜索算法,寻找到一个评分最高的网络结构。通常采用启发式的搜索算法,常用的方法有爬山法,模拟退火法,演化法和抽样算法。

搜索算法的原理就是对随机给定的一个贝叶斯网络结构做出有向边的修改,包括添加、删除、反向,并且保证每一步修正过后的网络结构评分高于前一步的网络评分,直到无法寻找到评分更高的网络为止。

## 4 交通事件数据集的分类和预处理

数据来源于荷兰的中部城市Utrecht,记录了Utrecht从2005年5月1日~9月13日的1 853个交通事件,主要来自于荷兰的国家事件管理中心,还有一部分来自交通相关部门的处理信息。每一组交通事件都包含了17个属性变量和一个类别变量。

为了提高预测精度,对事件进行分类处理,包括:① 交通事故小汽车有伤亡;② 交通事故小汽车无伤亡;③ 小汽车车辆抛锚;④ 卡车车辆抛锚;⑤ 货物掉落。

利用SPSS软件分别对5类事件进行显著性影响因子提取,分别得到相应类型事件的贝叶斯网络节点。

下面还要对交通事件数据中时间数据进行离散化处理。通常对数据离散化的处理,是对连续型的数值数据进行分段处理,将变量的取值分为几个区域。针对本文所使用的交通事件数据集,将时间节点离散化。根据相关研究标准,如果预测的误差的绝对值小于等于15 min,可视为有效预测。基于此,以15 min为一个区间对时间数据进行离散化,由于大于90 min的事件实例样本数过少,因此考虑到实际情况,对时间离散化如表1所示。

**表1 对时间的离散化**

**Tab.1 Discretization of time**

| 时间区域/min | 离散化分类 |
|---|---|
| [0~15) | 1 |
| [15~30) | 2 |
| [30~45) | 3 |
| [45~60) | 4 |
| [60~75) | 5 |
| [75~90) | 6 |
| ≥90 | 7 |

## 5 WEKA实验平台上的模型预测结果及评价

### 5.1 WEKA实验平台上的模型预测结果

对交通事件进行分类预测的工作在WEKA平台上完成。WEKA是一款全免费的开源的数据挖掘分析软件,并且给出了相当全面的数据挖掘分析算法,其中功能包括了数据预处理、分类、聚类、回归等[8]。使用

中国知网 https://www.cnki.net

WEKA对上述5大类事件数据分别进行预测，结果见表2～表6，分别有：交通事故小汽车有伤亡，交通事故小汽车无伤亡，车辆抛锚小汽车，车辆抛锚卡车，货物掉落。在5大类事件的数据集中，采用了80%的数据作为训练集来建立贝叶斯网络模型，选取其中20%作为测试集，来检测构建的贝叶斯网络模型的预测效果。

通常是以预测值与实际值的误差的绝对值在15 min之内，视为预测准确。根据这一标准，在评价预测结果的时候，选取这样一种评价方式：当前预测实例的实际值所在区间，以及该区间的前后2个区间，共3个区间内，若预测的分类结果落在这3个区间内，视为有效预测。

**表2　交通事故小汽车有伤亡预测分析**

**Tab.2　Prediction analysis of car accident casualties**

| 时间段/min | | 测试集实例数/个 | 有效预测数/个 | 有效率/% |
|---|---|---|---|---|
| 0~30 | 0~15 | 0 | 0 | 70 |
| | 15~30 | 10 | 7 | |
| 30~60 | 30~45 | 11 | 9 | 71.4 |
| | 45~60 | 17 | 11 | |
| >60 | 60~75 | 12 | 11 | 72.7 |
| | 75~90 | 6 | 4 | |
| | >90 | 4 | 1 | |
| 总测试集 | | 60 | 43 | 71.7 |

**表3　交通事故小汽车无伤亡预测分析**

**Tab.3　Prediction analysis of no car accident casualties**

| 时间段/min | | 测试集实例数/个 | 有效预测数/个 | 有效率/% |
|---|---|---|---|---|
| 0~30 | 0~15 | 18 | 14 | 76.1 |
| | 15~30 | 28 | 21 | |
| 30~60 | 30~45 | 20 | 17 | 75.8 |
| | 45~60 | 13 | 8 | |
| >60 | 60~75 | 4 | 2 | 40 |
| | 75~90 | 0 | 0 | |
| | >90 | 1 | 0 | |
| 总测试集 | | 84 | 62 | 73.8 |

**表4　车辆抛锚小汽车预测分析**

**Tab.4　Predictive analysis of car vehicle breakdown**

| 时间段/min | | 测试集实例数/个 | 有效预测数/个 | 有效率/% |
|---|---|---|---|---|
| 0~30 | 0~15 | 22 | 19 | 86.3 |
| | 15~30 | 29 | 25 | |
| 30~60 | 30~45 | 11 | 10 | 63.2 |
| | 45~60 | 8 | 2 | |
| >60 | 60~75 | 7 | 2 | 57.1 |
| | 75~90 | 3 | 1 | |
| | >90 | 3 | 1 | |
| 总测试集 | | 83 | 60 | 72.3 |

中国知网　https://www.cnki.net

表5 车辆抛锚卡车预测分析

Tab.5 Predictive analysis of truck vehicle breakdown

| 时间段/min | | 测试集实例数/个 | 有效预测数/个 | 有效率/% |
|---|---|---|---|---|
| 0~30 | 0~15 | 3 | 1 | 66.7 |
| | 15~30 | 3 | 3 | |
| 30~60 | 30~45 | 17 | 10 | 55.6 |
| | 45~60 | 10 | 5 | |
| >60 | 60~75 | 18 | 7 | 58.5 |
| | 75~90 | 6 | 5 | |
| | >90 | 17 | 12 | |
| 总测试集 | | 74 | 43 | 58.1 |

表6 货物掉落预测分析

Tab.6 Predictive analysis of cargo drop

| 时间段/min | | 测试集实例数/个 | 有效预测数/个 | 有效率/% |
|---|---|---|---|---|
| 0~30 | 0~15 | 28 | 25 | 94.4 |
| | 15~30 | 26 | 26 | |
| 30~60 | 30~45 | 15 | 15 | 95 |
| | 45~60 | 5 | 4 | |
| >60m | 60~75 | 4 | 1 | 22.2 |
| | 75~90 | 2 | 1 | |
| | >90 | 3 | 0 | |
| 总测试集 | | 83 | 72 | 86.7 |

### 5.2 预测结果评价

分析以上5大类事件的预测结果，可以看出除了卡车抛锚类型的事件，其他4类事件在总测试集的预测准确率上均超过了70%，而货物掉落类型的事件总测试集的预测结果甚至到达86.7%。

在低、中时间段(<30 min，30~60 min)贝叶斯网络模型的预测效果都达到了比较高的预测精确度，① 由于低、中时间段的事件实例数较多，机器对数据进行充分的学习，使预测达到较高精度；② 对于低、中时间段，事件持续时间相对较短，造成的误差也相对较小，预测精度也相对较高。以上5大类事件的预测结果中，货物掉落类型事件，在低、中时段的预测准确率在90%以上。交通事故小汽车无伤亡、有伤亡，在低、中时段的预测结果也都在70%以上，车辆抛锚事件由于本身的数据集的实例数比较少，因此预测精度稍微偏低。

而高时段的事件预测结果普遍不高，① 由于高时段的事件实例数并不多，机器无法对数据进行充分的学习，导致影响预测效果；② 交通事件本身就是一个随机性很大的问题，与事件处理人员的能力，事件发生时的天气等诸多因素有关系，也与事件数据采集本身相关，因此数据本身也可能有比较大的差异；③ 对于60 min以上的事件，时间本身比较长，那么对于此类事件来说，误差更大一些也通常可以接受。例如对于80 min以上的事件，那么误差超过20 min也是可以接受的。

## 6 结束语

交通事件持续时间的预测对交通管理者和出行者都有重要意义。首先介绍了贝叶斯网络方法的基本概念和构造贝叶斯网络的要素，以及贝叶斯网络预测模型的优势，提出了对交通事件采取分类的处理方法。通过对荷兰交通部门提供的交通事件的分析，选择出事件中的显著性变量，利用数据挖掘软件WEKA，对事件数据进行相关处理，建立完整的贝叶斯网络模型。分析其预测结果表明：贝叶斯网络模型在预测交通事件持续时间的工作上，有较好的精度，值得进一步研究和挖掘。

中国知网 https://www.cnki.net

此外,贝叶斯网络模型对提高交通事故预测准确性应注意的问题:① 贝叶斯网络的概率学习是一个比较繁琐的过程,应当获取更多的交通事故事件的实例数,来提高预测的精度;② 交通事件不但本身随机性很大的,而且是多个因素综合作用的结果,在建立预测模型的时,应尽量考虑更多的因素,来提高预测的准确性。

**参考文献:**

[1] SCHRANK D,LOMAX T. The 2004 urban mobility report[J].Texas Transportation Institute's Annual Urban Mobility Report,2004,9(1):27-31.
[2] 姬杨蓓蓓.交通事故持续时间预测方法研究[D].上海:同济大学交通运输工程学院,2008.
[3] 康国祥,方守恩.Cox Regression模型在交通事件持续时间研究中的应用[J].交通信息与安全,2011,2(29):104-106.
[4] 刘伟铭,管丽萍,尹湘源.基于多元回归分析的事件持续时间预测[J].公路交通科技,2005,11(22):126-129.
[5] 孟祥海,郑来,秦观明.基于模糊逻辑的交通事故预测及影响因素分析[J].交通运输系统工程与信息,2009,9(2):87-92.
[6] 张连文,郭海鹏.贝叶斯网引论[M].北京:科学出版社, 2006:39.
[7] 胡春玲.贝叶斯网络的结构学习算法研究[D].合肥:合肥工业大学,2006.
[8] VAPNIK V N. Anover view of statistic all earning theory[J]. Trans Neural Netw,1999,10(3):988-999.
[9] 周雪峰,郑长江.基于博弈论的无控制路段人行横道处人车抢行分析[J].华东交通大学学报,2012,29(6):65-69.

# Traffic Incident Duration Prediction Based on Bayesian Network

Zheng Changjiang, Ge Shengyang, Zheng Shukang

(College of Civil and Transportation Engineering, Hohai University, Nanjing 210098, China)

**Abstract:** With the continuous improvement of data collection instruments and related research and technological development, establishing models based on data mining has become the main direction for studying the traffic incident duration. Based on traffic incident data from the Dutch transport sector, this paper conducts classification and pre-processing, analyzes the event duration frequency chart, and classifies the collected data according to the typical event category. By using principal component analysis and stepwise regression to extract significant impact factor, it establishes Bayesian network model through data mining software WEKA .Then with 80% of the data in the dataset to learn modeling, 20% of the data as a test set to test the predicted effects, this study makes performance evaluation. Experimental results show that compared to other prediction methods, Bayesian network model algorithm has higher prediction accuracy and high randomicity for a number of large traffic events with many variables.

**Key words:** urban traffic; traffic incident duration; Bayesian network model; datasets classification; impact factor extraction; WEKA

中国知网　https://www.cnki.net