

# 基于大数据分析的实时交通安全监测与预警研究<sup>①</sup>

魏传佳<sup>1</sup> 何世伟<sup>2</sup>

<sup>1</sup> (泉州轻工职业学院 智能工学院, 福建 泉州 362200)

<sup>2</sup> (台湾大学计算机系, 台湾 10617)

**摘要** [目的] 目标是通过减少拥堵和碰撞风险来改进和提高城市交通系统的性能。[方法] 通过数据挖掘和贝叶斯推理技术在实时碰撞预测模型中的使用, 确定影响交通运行的因素, 包括间接拥堵位置和直接拥堵位置。[结果] 实验仿真结果表明共同监测、改善交通运营和安全的重要性。[局限] 对于超大规模的高速模型预测有待后续改进算法仿真其效率。[结论] 该技术能有效避免交通拥堵现象的发生, 同时保证交通道路的有效利用率。

**关键词** 大数据; 实时; 安全性; 数据挖掘

## Research on real-time traffic safety monitoring and early warning based on big data analysis

WEI Chuanjia<sup>1</sup> HE Shiwei<sup>2</sup>

<sup>1</sup> (Department of Artificial Intelligence Management, Quanzhou College of Technology, Quanzhou 362200, China)

<sup>2</sup> (Department of Computer Management, National Taiwan University 10617, China)

**Abstract** [Objective] The goal is to improve the performance of urban transport systems by reducing the risk of congestion and collisions. [Methods] In this paper, data mining and Bayesian inference techniques are used in real-time collision prediction models to determine the factors affecting traffic operations, including indirect (peak hours, quantity and upstream low speed) congestion locations and direct congestion locations (higher

①本文系2018年度福建省中青年教師科研項目(高校教育信息化專項)(項目編號: JZ180209)的研究成果之一。

downstream congestion index) . **[Results]** Experimental simulation results demonstrate the importance of joint monitoring and improved traffic operations and safety. **[Limitations]** For ultra-large-scale high-speed model prediction, it is necessary to improve the algorithm and simulate its efficiency. **[Conclusions]** This technology can effectively avoid the occurrence of traffic congestion and ensure the effective utilization of traffic roads.

**Keywords** Big Data; Data real-time; Security; Data Mining

## 1 引言

在过去的几十年中, 由于智能交通<sup>[1]</sup>系统 (ITS) 的快速普及使得交通领域的大数据在不同的地理范围和不同来源得以收集。海量的看似无序的数据, 可以大大增强专家系统地理理解它们。此外, 由于运输中大数据的实时性, 因此可以实现主动流量管理, 使得交通系统的性能得以提高。长期以来, 运营效率和交通安全一直被视为公路系统绩效评估的重中之重。效率可以用交通拥塞来衡量, 而安全则是通过碰撞分析来研究的。

本文以福建中南部高速公路管理局的系统为研究对象。这个系统由三条高速公路组成,

这些高速公路位于人口密集的城市地区。收费高速公路是相通的市区, 机场和其他地方, 该系统设有多个ITS系统, 用于电子收费和记录旅客信息。275个MVDS探测器分布在高速公路上, 平均间距小于1500米, 综合交通流参数按1分钟为准, 因此, 大规模的地理部署和连续的数据收集为全面的网络性能评估提供了可靠的大数据来源。本文基于MVDS数据进行实时运行和安全分析, 期望大数据分析对实时交通运行和安全监控起到优化的作用。

### 准备数据

本研究中由G72、G15和G76组成的高速公路系统, 如图1所示。G72穿越德化县、永春县、泉州市, 可以容纳更多的通勤。G15连



图1 福建中南部高速地图

接泉州、漳州市。G76连接龙岩市、漳州市。G76和G72两者都与G15连接。在整个系统上，总共安装了275个MVDS探测器。如表1所示，目前在每条高速公路上部署的MVDS系统得到了很好的覆盖，相邻探测器之间的平均距离小于1.5千米。

研究表明，2016—2018期间三条高速公路共发生581起事故，其中243条为后端拥塞，

如表2所示。与其他两条高速公路相比，G15在三条高速公路中具有最大的碰撞数和最高后端碰撞率。这种现象归因于G15上的大量交通量和通勤量。

对于每次碰撞，碰撞前5—10分钟的交通数据来自两个上游和两个下游MVDS探测器，如图2所示，最接近碰撞位置的数据被收集和累加。

表 1 高速公路 MVDS 系统分布

高速路名称	长度 (千米)	方向	探测器个数	相邻探测器距离			
				平均值	标准差	最小值	最大值
G72	22.4	北	26	0.83	0.78	0.09	2.99
		南	29	0.83	0.81	0.09	3.00
G76	21.4	北	55	0.38	0.18	0.10	1.00
		南	55	0.18	0.18	0.10	1.00
G15	31.5	东	55	0.53	0.27	0.2	1.30
		西	55	0.53	0.27	0.2	1.20

表 2 高速公路后端拥塞统计

高速路名称	后端拥塞		
	不是	是	总计
G72	99	53	152
G76	106	33	139
G15	133	157	290
总计	338	243	581

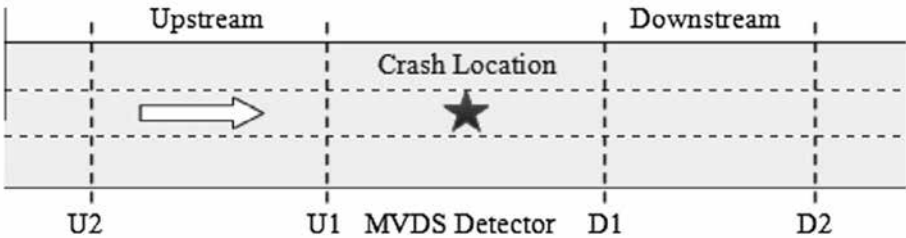


图 2 拥塞区与侦测器位置

## 2 算法设计

### 2.1 实时拥塞监控

本文采用基于速度的拥塞指数来测量空间和时间尺度上的拥塞强度。自由流速是检测第 85 个百分位速度位置。根据公式 (1), CI 是

$$\text{拥塞指数 (CI)} = \frac{\text{自由流速度} - \text{实际速度}}{\text{自由流速度}} \\ = 0$$

### 2.2 随机森林

随机森林<sup>[2-4]</sup>是一个集合分类器, 使用许多决策树模型选出最受欢迎的类, 单个决策树会产生高差异或偏差。相比之下, 随机森林提供了无偏估计。此外, 强大的大数定律保证了随机森林对过度拟合的鲁棒性。

在实时交通安全评估中, 随机森林的基本应用是评价变量的重要性。当该单个变量的 OOB (包外) 数据被置换时, 随机森林算法通过查看预测误差增加多少 (或精度降低多少) 来评价变量的重要性。另一个衡量标准是用节点总减少量的基尼系数来表示, 该变量系数来自所有树的平均值。前一项评价的缺点是高估了相关变量的重要性。后一项对许多类的因子预测表现不佳。

### 2.3 贝叶斯 logit 模型

为了预测实时拥塞的可能性, 评估了贝叶斯框架下的逻辑回归模型<sup>[5-7]</sup>。逻辑回归模型及其扩展已被广泛用于实时安全性研究, 其中包含来自不同来源的数据, 几何特征和天气数据也被证明在逻辑模型中很有用。因此, 这种统计方法能够处理来自不同来源的信息。随着大数据的快速发展, 预计未来可以将新数据来源纳入此建模框架中。目标变量是拥塞发生的二进制指标, 拥塞情况的概率为  $p (y = 1)$ , 非拥塞情况的概率为  $1 - p (y = 0)$ 。本文构

一个在 0 和 1 之间的连续变量。CI 值的增加表明拥塞程度增加, 在每个检测位置上, CI 以 5 分钟的间隔完成聚合, 通过创建填充等高线图显示其拥塞分布。

当  $CI > 0$  时, (1)

当  $CI < 0$  时,

建了三种逻辑模型, 并对它们的性能进行了比较: I 匹配事件对照逻辑模型; II 固定效应逻辑模型; III 区分峰值和非峰值时间随机参数逻辑模型。公式如下:

$$y = \text{二项式} (p_i, 1) \quad (2)$$

对于模型 I,

$$\text{logit} (p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = X_i \beta + \varepsilon_{j|i} \quad (3)$$

对于模型 II,

$$\text{logit} (p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + X_i \beta \quad (4)$$

对于模型 III,

$$\text{logit} (p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = \beta_{0|t} + X_i \beta_t \quad (5)$$

其中  $\beta_0$  为常数项,  $\beta$  是解释变量系数的向量。除了用于峰值 ( $t = 1$ ) 和非峰值 ( $t = 0$ ) 交通时间外,  $\beta_{0|t}$  和  $\beta_t$  具有相同的含义。 $\varepsilon_{j|i}$  表示对第  $j$  组内所有项的 logit 贡献。在贝叶斯推理中, 首先证明参数的预先分布是合理的, 本推理设置为无信息先验。对于  $\beta_0/\beta_{0|t}$  和  $\beta/\beta_t$  中的每个元素, 取值范围  $(0, 10^6)$ ,  $\varepsilon_j$  具有正态分布  $(0, 1/\tau)$ , 其中  $\tau \in (0, 0.001)$ 。

使用 WinBUGS 软件进行模型校准。模拟三条链路, 在 15000 次迭代中, 将前 5000 次迭代作退化模拟。通过检查三条链路的迹线图是否彼此重叠, 来确保参数的收敛性。本文使用偏差信息准则 (DIC) 作为一种评价贝叶斯模型复杂度和拟合度的方法, 较小的 DIC 预

示着更好的数学模型。使用贝叶斯可信区间 (BCI) 用于参数估计, 假设95%的 (BCI) 不包含0, 那么变量的影响是非常显著的。

### 3 拥塞监控与建模结果分析

#### 3.1 拥塞评估

如上所述, 每个站点以5分钟的间隔聚合CI值。为了获取更稳定拥塞段和持续时间的数值, 本文以每天的数据进行持续计算。TTI表示CFX的拥塞度, 1.25和2.0的TTI定义为中度和高度拥塞的阈值。给定两个TTI拥塞阈值的

实际行进速度和自由流速度之间的比率是4:5和2:1, 它们分别相当0.2和0.5的CI值。所以, 设定CI值为0.2和0.5作为中等和高拥塞阈值<sup>[8]</sup>。如图3(a—d)所示, 拥塞程度与特定时间和当前拥塞条件关系密切。对于同一条高速公路, 在多个检测点上识别早晨和晚上的高峰时间。对于相同的位置, 拥塞程度在一天中的不同时间有着显著的不同。因此, 为了实现更准确的拥塞检测, 需要连续监视。对于交通安全研究, 还应使用实时拥塞测量来揭示拥塞对碰撞事件的真实影响。

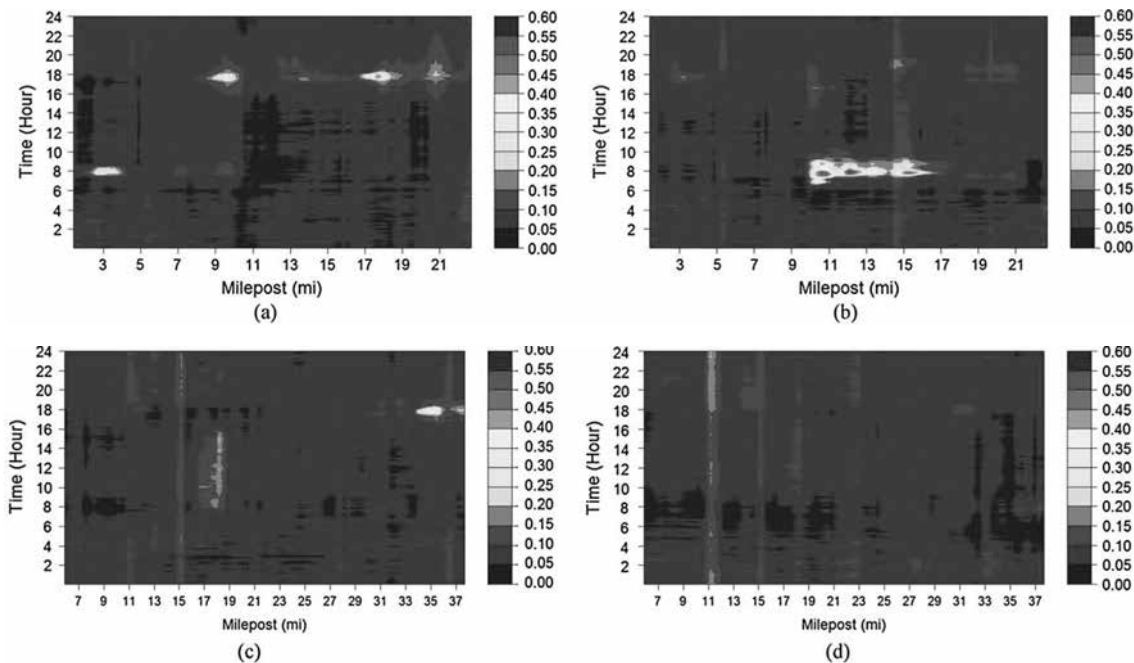


图3 (a—d) 拥塞程度仿真关系图

#### 3.2 变量选择

基于R的数据挖掘工具Rattle构建了变量选择的随机森林模型。在模型中, 每次拆分时, 随机抽取4个变量。在随机森林中树的数量不应太小, 以确保每次模拟至少可以预测几次, 本算法设置了500个树。从37个变量中找出重要性排名前20的变量, 剔除其他不重要

的17个变量。图4描述了两种不同的变量排序方法, 但是得出的前20个重要变量的类型差异不大。数量, 高峰时段, 平均速度和拥塞指数的对数是关键变量。然而, 在确定要纳入最终模型的变量之前, 应检验变量之间的相关性。为解决此问题, 作者在表3中进行了相关性检验, 并进行了简单的逻辑回归分析, 除了控制



相关性之外还保留重要变量。

综合随机森林, 相关性检验和初步Logistic

回归的结果, 得到4个变量。表4中提供了这些变量的描述性统计数据。

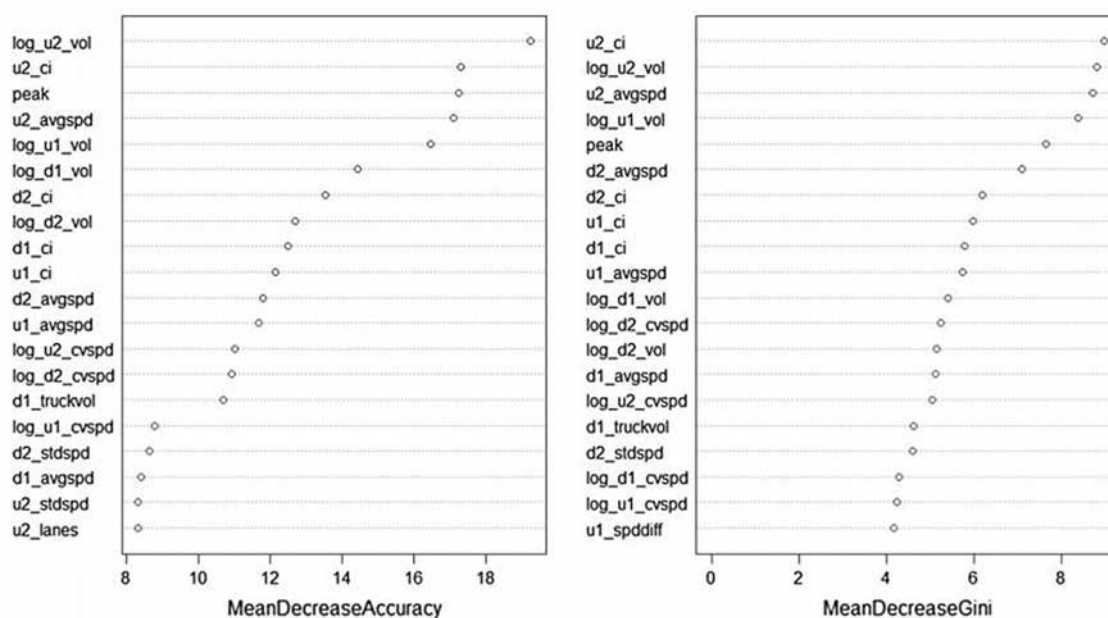


图 4 基于随机森林的变量重要性曲线

表 3 变量模型中的相关性测试结果

Pearson's correlation	peak	log_u2_vol	u2_avgspd	d1_ci
peak	1.0000	0.3437	-0.2767	0.3289
log_u2_vol	0.3437	1.0000	-0.2426	0.1331
u2_avgspd	-0.2767	-0.2426	1.0000	-0.3705
d1_ci	0.3289	0.1331	-0.3705	1.0000

表 4 变量模型中统计分析结果

Description		Mean	Std Dev	Min	Max
crash	Rear-end crash: non-crash=0; crash=1	0.201	0.401	0.000	1.000
peak	Peak hours: nonpeak=0; peak=1	0.161	0.368	0.000	1.000
log_u2_vol	Log volume of U2 station	4.611	1.118	0.693	6.762
u2_avgspd	Average speed of U2 station	62.138	8.735	2.500	98.000
d1_ci	Congestion index of D1 station	0.058	0.103	0.000	0.909

## 4 结论

智能交通系统在过去几十年的快速发展促进了大数据在交通领域的实施。利用大数据的强大功能来提高流量系统的性能。本研究中,使用实时微波车辆检测系统(MVDS)进行数据监测和改善福建中部城市高速公路交通运行。从体积、速度和变化的角度来看,MVDS应被视为大数据的获取的主要来源。检测系统按每分钟每车道的车辆类型获取实时的速度、体积、车道占用率。基于这些数据,为3条高速公路开发了拥塞检测和实时安全分析系统。

传统解决拥塞的措施缺乏捕捉拥塞变化的能力。因此,更期望基于大数据的实时拥塞测量以识别时间和空间中的拥塞模式。引入拥塞指数来测量拥塞强度并通过填充等高线图来实现可视化。研究发现城市高速公路的拥塞时间和地点都很高。在特定位置观察到早晨和晚上高峰时段的复发性拥塞。面对高峰时段的大量交通需求,交通管理部门不能总是扩大系统容量作为解决方案。目前,DMS已被广泛应用于旅行时间估算的CFX系统。但是,它也可用于拥塞警告。拥塞地点和潜在延误的信息将使驾驶员有足够的时间来调整他们的速度并提高他们对周围交通的防范意识。如果实现更平滑的交通流量,则预计将减轻拥塞。数学分析和仿真结果表明,该技术能有效避免交通拥堵现象的发生,同时保证交通道路的有效利用率,为交通管理部门和广大行人及车辆操作人员的出行提供了科学、可行的参考。

## 参考文献

- [1] 罗东健. 大规模存储系统高可靠性关键技术研究 [D]. 华中科技大学, 2011.
- [2] 王健宗. 云存储服务质量的若干关键问题研究 [D]. 华中科技大学, 2012.
- [3] 陈勇, 黄席樾, 杨尚昱. 汽车防撞预警系统的研究与发展 [J]. 计算机仿真, 2006, 23 (12): 36—38.
- [4] 王艳军, 吕志勇, 黄蕾. 基于物联网传感器的城市交通状态预测 [J]. 武汉理工大学学报, 2010, 32 (20): 108—111.
- [5] Bianca Schroeder, Garth A. Gibson. Understanding disk failure rates [J]. ACM Transactions on Storage (TOS), 2014 (3): 45—50.
- [6] Lakshmi N. Bairavasundaram, Garth R. Goodson, Shankar Pasupathy, Jiri Schindler. An analysis of latent sector errors in disk drives [J]. ACM SIGMETRICS Performance Evaluation Review, 2015 (1): 78—85.
- [7] Ahmed, M.M., Abdel-Aty, et al. The viability of using automatic vehicle identification data for real-time crash prediction [J]. IEEE Trans. Intell. Transp. Syst, 2012, 13 (2): 459—468.
- [8] Z Xue Gang, L Wei. Research in the Model of the Area Road Traffic Security Risk Assessment [C]. Chongfu Huang, XilinLiu. Theory and Practice of Risk Analysis and Crisis Response. Amsterdam - Paris: Atlantis Press, 2008. 857—861.