文章编号：0253-374X(2015)02-0221-05

DOI:10.11908/j.issn.0253-374x.2015.02.009

快速路交通流运行安全关键参数识别与评估

贾丰源[1]，孙　杰[2]，孙　剑[2]

(1. 同济大学 汽车学院，上海 201804；2. 同济大学 道路与交通工程教育部重点实验室，上海 201804)

摘要：基于上海市两条快速路采集的事故数据和相应检测器数据，应用随机森林模型对事故发生前 5～10 min 内的交通流数据进行重要变量筛选. 利用基于高斯混合模型和最大期望算法的贝叶斯网络(BN)模型对快速路实时交通流事故风险进行建模分析，并对建立的 BN 模型进行了可转移性测试. 结果表明：选取重要变量后建立的 BN 模型效果优于使用直接检测数据建立的模型，事故预测准确率达到 82.78%；可转移性测试中 BN 模型的事故预测准确率虽有所下降，但整体预测精度和事故预测精度仍都优于利用直接检测数据建立的模型.

关键词：城市快速路；交通安全；主动风险评估；检测器数据；随机森林；贝叶斯网络(BN)

中图分类号：U121　　**文献标志码**：A

Key Variables Identification and Proactive Assessment of Real-time Traffic Flow Accident Risk on Urban Expressway

JIA Fengyuan[1], *SUN Jie*[2], *SUN Jian*[2]

(1. School of Automotive Studies, Tongji University, Shanghai 201804, China; 2. Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China)

Abstract: Based on accident data and detector data collected on two expressways in Shanghai, important variables for model construction were selected from the data of traffic flow within 5～10 min before the accident with random forest model. Then, the Bayesian network (BN) model based on the Gaussian mixture model and expected maximum algorithm was established for the analysis of real-time traffic flow state and accident risk. Meanwhile, the transferability of BN model was also assessed. The results show that BN model built with selected important variables is better than that with direct detection data, with the accident prediction accuracy rate of 82.78%. The results of the transferability show that the improved BN model is still better than the traditional model, though the accident prediction accuracy of BN model decreases.

Key words: urban expressway; traffic safety; proactive assessment of accident risk; detector data; random forest; Bayesian networks (BN)

城市快速路偶发性交通事故不仅会造成人员及财产损失，同时亦是交通拥堵的重要致因. 尤其是在高峰时段，虽然事故严重程度低，但是由此引起的交通拥堵和延误会极大地浪费社会成本.

导致交通事故发生的原因主要可分为人、车、路和环境四个方面，而快速路的实时交通流运行状态变化是四个方面综合作用的外在表现. 近年来，利用城市快速路实时交通流检测数据主动估计交通流运行风险得到了广泛关注. 其方法是利用有/无事故发生时的事件数据和相对应的路段上下游交通流检测器数据建立合适的模型，以预测道路上发生事故的风险.

快速路的交通流运行风险可由很多变量及其组合描述，常用的包括快速路检测器采集到的基本交通流参数(流量、速度和占有率)及其组合、道路线形参数及环境参数等. 目前的研究主要集中于模型方法的选取，而忽视了在较多能描述交通流状态的变量中筛选影响安全风险的关键变量. 使用较多变量进行建模不仅容易导致模型具有较高的计算复杂度，而且容易出现过拟合情况.

本文针对上海市快速路事故视频监控系统以及事故救援报警信息整理的快速路事故数据，选择 7 处道路线形和环境参数相似的路段，通过随机森林(random forest, RF)模型选取事故发生前 5～10

收稿日期：2014-04-15

基金项目：教育部新世纪人才计划(NCET-13-0425)；中央高校基本科研业务费专项资金(1600219205)

第一作者：贾丰源(1983—)，男，博士生，主要研究方向为交通安全、汽车被动安全. E-mail:gilbert1000@163.com

通讯作者：孙　剑(1979—)，男，教授，博士生导师，工学博士，主要研究方向为交通仿真与实验、交通流分析与优化. E-mail:sunjian@tongji.edu.cn

min内重要的交通流状态数据,再使用改进的贝叶斯网络(Bayesian networks, BN)模型对快速路的事故风险进行实时预估,并对建立的模型进行可转移性测试,即测试通过训练数据建模的预测模型能否用于其他快速路上.

1 研究综述

目前,国外针对高速路的事故风险预测做了很多研究,利用不同的建模方法对事故风险进行分析和预测,包括不同的统计回归方法[1-5]和数据挖掘算法[6-15]等. 其中,Oh等[1]最早使用非参数贝叶斯统计方法对快速路上的实时交通流运行风险进行预测,Lee等[2]使用集计对数线性模型确定可能会导致事故发生的交通流状态,Abdel-Aty等[3-4,6-9]也在实时事故风险评估方面做了一系列的工作.

在最近的研究中,Pande等[10]使用美国的四条高速公路数据以及逻辑回归和分类树二元分类方法建立了实时事故风险评估模型,模型的可转移性在此次研究中也得到测试;Hossain等[11]以及Pham等[12]也分别基于日本和瑞士的快速路事故数据建立了实时的事故预测模型;Ahmed等[5,13]根据从自动车辆识别系统(AVI)中提取的交通流数据对实时的事故风险进行了分析.

对于模型建立前的关键变量筛选,在数据建模领域已有了诸多研究. 在入侵检测方向,Zainal等[16]在建立入侵检测模型前使用粗糙集(rough set)识别重要属性. 在事故风险预测领域,Hossain等[11,14]也曾使用随机多项分对数模型(RMNL)筛选重要变量后建模.

综上所述,已有不同的建模方法用于国外高/快速路实时事故分析预测,而我国的城市快速路几何特征、交通流特征及事故特征均与国外有所差异. BN模型可用于预测实时的事故风险,且效果较其他几种事故预测算法更好,但未进行影响模型精度的重要变量筛选,且未进行可转移性测试[15]. 因此,本研究在建立BN模型前使用RF模型判断影响变量的重要程度,再对选取的重要变量进行建模,以达到简化模型输入并提高模型预测精度的目的,并且使模型具有较高的可转移性.

2 研究路段和数据

2.1 研究路段

本文主要研究快速路主线基本段交通运行状态对事故风险的影响. 考虑到快速路的线形以及入口和出口匝道的影响,本文选取了上海市两条快速路上共7段线形和匝道类型基本一致的主线路段,包括延安高架4段和南北高架3段. 延安高架采集的数据用于建模和评价,而南北高架的数据用于测试模型的可转移性. 本文所使用的7路段均为3车道,路段上线圈检测器的间隔为300～500 m. 各路段名称与编号对应关系见表1.

2.2 数据

本文研究的事故数据由事故视频监控系统中提取. 由于研究需要将每条事故信息和相应地点与时间的交通流状态信息对应起来,同时由于事故的偶发性特征,因此在路段上获取大量的事故数据异常困难. 经过交通流检测数据和事故数据的质量控制和无效信息剔除后,目前采用了2010年7月研究路段上发生的90条两车相撞或多车相撞的事故数据,其中延安高架上71条,南北高架上19条. 每条事故数据的信息包括事故的发生时间、事故地点、事故类型、事故发生时的天气和位置等信息,不同地点的事故数和所占比例见表1.

本文使用的快速路交通流状态由事故地点附近的线圈检测器采集到的基本交通流数据(流量、速度和占有率)及其组合计算的结果表示. 由前期研究结果可知[15],使用事故发生地点上下游各一个检测器在事故发生前5～10 min内的交通流数据建立的

表1 事故数据统计表

Tab.1 Summary of crash data

快速路	事故地点	编号	事故数	所占比例%
延安高架西向	延东立交入口匝道至茂名路上匝道	a	23	25.6
延安高架东向	虹许路上匝道至娄山关路上匝道	b	17	18.9
	延西立交入口匝道至凯旋路上匝道	c	11	12.2
	江苏路上匝道至华山路上匝道	d	20	22.2
南北高架北向	延长路下匝道至广中路上匝道	e	5	5.6
南北高架南向	广中路下匝道至洛川路上匝道后段	f	11	12.2
	共江路上匝道至场中路上匝道	g	3	3.3
总计			90	100

BN 模型效果最好，故模型选取事故发生前 5～10 min 内对应的交通流状态数据建立模型. 由于每条事故数据对应的交通信息只来自于事故发生地附近的四个检测器——事故上下游最近的检测器各两个. 可将四个检测器按由上游至下游的顺序分别命名成 D1，D2，D3，D4，如图 1 所示. 因此，根据这四个检测器的原始数据，可分别计算事故地点上下游各一个检测器的流量、速度和占有率的差值及上下游的平均值.

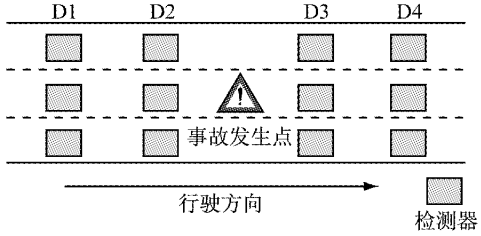

图 1　事故发生点检测器示意图

Fig.1　Arrangement of detectors at the crash site

模型的输入变量可分为三部分：第 1 部分为检测器的原始变量；第 2 部分为上下游各检测器变量的差值；第 3 部分为上下游各检测器变量的平均值，具体变量可见表 2. 其中数字 1～4 分别对应检测器 D1～D4，q_1，v_1，o_1 分别表示检测器 D1 在 5 min 内的流量、平均速度和平均时间占有率，而 D_q_13 表示 D1 和 D3 两个检测器的流量差，A_q_12 表示 D1 和 D2 两个检测器的平均流量，其他变量命名类似.

表 2　可选变量列表

Tab.2　Candidate variables for modeling

可选变量	变量命名
检测器的原始变量	q_1，v_1，o_1，q_2，v_2，o_2，q_3，v_3，o_3，q_4，v_4，o_4
上下游各检测器变量的差值	D_q_13，D_v_13，D_o_13，D_q_14，D_v_14，D_o_14，D_q_23，D_v_23，D_o_23，D_q_24，D_v_24，D_o_24
上下游各检测器变量的平均值	A_q_12，A_v_12，A_o_12，A_q_34，A_v_34，A_o_34

为了能够基本表现实际事故发生与否的交通状态比例，同时满足本方法建模的数据需求，因此随机生成了非事故数据 904 条，使得事故数据与非事故数据的比例接近 1∶10.

3　模型构建

3.1　随机森林模型

如表 2 所示，共有 30 个可选变量用于模型建立. 考虑到模型的复杂性，在建立模型前，需要筛选出对快速路交通流安全风险影响较大的变量作为模型的输入变量. RF 是 Breiman 于 2001 年提出的一种集成学习方法，可以有效地计算出模型变量的重要性[17]. 与其他传统变量选取方法如逻辑回归(logistic regression)不同，由于 RF 采取抽样法并且随机选取变量进行建模，所以可以较好地处理变量的多重共线性问题，并能够处理很高维度(很多变量)的数据，而且无需做特征选择，同时无需交叉验证来评估模型优劣[14].

构建 RF 模型的基本方法是：在原始训练样本集 L 中随机抽出 N 个样本作为一个新的样本集 L_b 来建立一个分类回归树(CART) T_b，并在这个树的每个节点处随机选择所有 p 变量中的 m 个变量进行分类树的节点分割. 重复上述步骤 B 次以形成一个 RF 模型.

使用 RF 模型来计算变量重要性时，则需要以下过程[18]：

(1) 确定每次生成的分类树 T_b 的袋外数据(OOB)，$L_{b,OOB}=L-L_b$.

(2) 使用生成的分类树 T_b 对 $L_{b,OOB}$ 进行预测分类，并计算分类正确的次数.

(3) 对每个变量 $j=1,\cdots,p$：① 变换 $L_{b,OOB}$ 中的变量值 x_j；② 使用 T_b 和变换的变量值 x_j 对 $L_{b,OOB}$ 进行预测分类，并计算分类正确的次数；③ 计算 OOB 数据变换变量后分类正确率的降低值.

最后所有 B 个分类树中 OOB 数据降低的平均准确率即变量 x_j 的重要性. 此外，需要注意的是，由于 RF 是随机抽取样本构建多个 CART，为减少 OOB 的形成，RF 模型构建的树的数量设为 1 000.

利用 Matlab 计算平台，实现 RF 识别程序. 将所有 30 个变量输入 RF 模型中进行计算测试，各变量重要性如图 2 所示. 其中，纵坐标表示该变量降低的平均模型准确率，以此表示其重要性.

前期研究表明，使用事故发生地点上下游各一个检测器在事故发生前 5～10 min 内的交通流数据建立的 BN 模型效果最好，该模型选取的变量数为 6 个[14]. 因此在本文中，为体现关键因素识别的作用，选择尽量少的输入变量个数，但又能保证模型预测准确率. 经过比较分析，最终本文选取重要性最高的 4 个变量作为模型的输入. 由图 2 可知，重要性最高的 4 个变量为 A_v_34，A_o_12，v_2，o_2，分别表示事故下游的平均速度、事故上游的平均时间占有率、D2 检测器的检测速度和时间占有率.

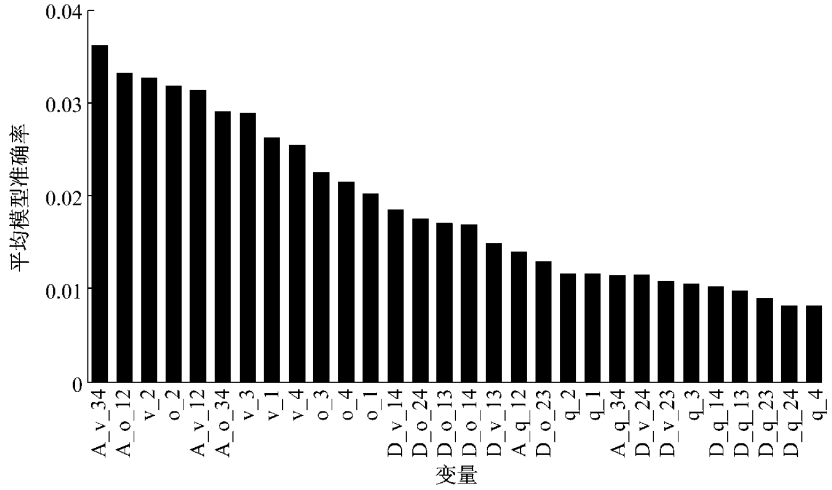

图2 变量重要性

Fig.2 Importance of variable

3.2 BN构建

在利用RF模型筛选出重要变量后，进一步的工作即是针对选取的变量建立实时的交通流运行安全风险预测模型. 快速路的安全风险问题存在着不确定性，即交通流在不同的流量、速度和占有率的组合下，它的安全风险是可变的、不确定的.

BN作为贝叶斯公式的扩展，可以处理人工智能研究中的不确定性问题. 建立BN主要目的是进行概率推理，用概率论处理不确定性以保证推理结果的正确性. 在交通事故分析、交通事件检测等方面，BN已经表现出优异的预测效果[19-20]. 在构造BN时，使用高斯混合模型描述输入的变量以及最大期望(EM)算法用于参数的学习可以有效应对缺失数据的情况，因此本文使用基于高斯混合模型和EM算法的BN模型对快速路实时交通流运行风险进行评估，而有关于BN模型的详细介绍在此就不再赘述，可详见文献[15].

本文基于Matlab实现了一个BN分类器，其中高斯混合模型和EM算法用于学习和训练BN的结构和参数. 整个BN分类器的构建分为以下三个阶段：

第1阶段是数据准备. 针对事故数据与非事故数据以及相应的交通流数据，使用随机的方法确定训练样本和测试样本. 为保证模型检验的可靠性，训练样本和测试样本具有相同的比例.

第2阶段是分类器的训练. 根据训练数据和基于高斯混合模型的EM算法来构建BN分类器.

第3阶段是分类器的应用. 使用分类器对测试样本进行分类，其输入是已经构建好的分类器的推断引擎和测试数据，输出是测试样本属于各类别的后验概率，最后根据这一概率进行分类.

在进行模型的可转移性测试时，使用已经训练好的分类器直接对另一条快速路上的数据进行测试，以得到模型的可转移性.

3.3 模型检验

本文分别使用事故数据的分类准确率、非事故数据的分类准确率以及整体的分类准确率三个标准来检验模型的有效性. 其中，事故数据的分类准确率为分类准确的事故数据数占总事故数据数的比例. 同样，非事故数据的分类准确率为分类准确的非事故数据数占总非事故数据数的比例. 考虑到事故的发生会带来严重的影响，本文的判断指标优先考虑模型对事故数据的分类率.

由于训练样本和测试样本是将同一数据样本随机分成比例一样的两个部分，所以每次针对同一整体数据样本建模得到的模型分类性能都不会完全一致. 因此，本文在进行建模时，对每类组合都分别测试10次，这样可以得到每组数据模型分类结果的平均值，以此来判断哪组数据建立的模型性能较优.

4 模型结果

4.1 BN模型结果

基于延安西路4个主线路段的71条事故和702条非事故数据，对事故发生前5～10 min内的交通流状态数据选取重要变量后进行建模分类，并与前期研究结果对比，结果见图3.

由图3可以看出，使用事故发生前5～10 min内的交通流状态数据建立的BN分类器的总体准确率和非事故准确率较高，与此同时也能得到较好的

事故分类准确率. 同时图 4 也显示，在选取 4 个重要变量后建立的 BN 模型，比原有研究中选取事故地点最近的两个检测器在事故发生前 5～10 min 内的检测数据建立的模型效果要好，不仅非事故分类准确率得到小幅度提升，事故分类准确率更提高到 82.78%. 因此，选取重要变量建立模型可以减少模型复杂度，而且可提高模型预测效果.

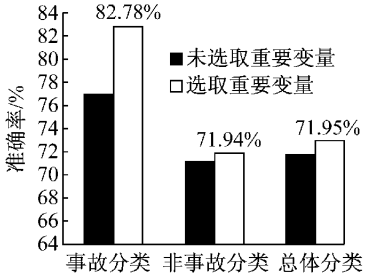

图 3　BN 模型预测准确率对比

Fig. 3　Comparison of prediction accuracy for BN model

4.2　BN 模型可转移性测试

本文也对 BN 模型的可转移性进行了测试. 本文中，使用延安高架上的数据建立的 BN 模型测试南北高架的数据. 基于南北高架上的 19 条事故数据和 202 条非事故数据，BN 模型的可转移性测试结果列于图 4 中. 与 BN 模型评估相同，本文对选取重要变量后与前期研究中只使用原始数据建立的模型的可转移性测试结果进行对比分析.

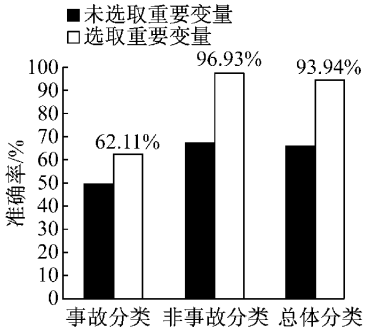

图 4　BN 模型可转移性测试

Fig. 4　Transferability of BN model

由图 4 可知，选取重要变量后建立的 BN 模型在测试南北高架的数据时也表现较好. 对比图 4 数据，基于可转移性测试的事故分类准确率有所下降，但非事故分类准确率得到很大提高，并使得总体分类准确率达到 93.94%. 与此相比，只使用检测器原始数据建立的 BN 模型得出的分类准确率就不尽如人意. 如 Pande 等[10]使用逻辑回归和分类树建立的事故预测模型分别可以分辨出 61.7%和 61.9%的事故，其模型的可转移性分别降为 55.84%和 50.37%.

5　结论

(1) 利用 RF 模型在事故发生前 5～10 min 内的事故地附近的四个检测器的原始检测数据和计算数据中选取 4 个最重要的变量，并基于高斯混合模型的 EM 算法构建 BN 模型. 与前期研究的结果对比表明，选取重要变量不仅能减少模型复杂度，也可以提高事故预测准确率，其可达到 82.78%.

(2) 对 BN 模型可转移性测试结果表明，使用一条快速路数据建立的模型可以用于其他快速路. 尽管事故的预测准确率有所下降，选取重要变量后的 BN 模型的可转移性仍优于未选取重要变量的模型. 另外，模型的预测精度和可转移性相较于前人研究都有一定优势.

进一步的工作是针对不同事故类型和不同车道事故特征的主动估计和预测分析，以进一步提高模型在实践中应用的稳定性和精度.

参考文献：

[1] Oh C, Oh J S, Ritchie S G, *et al*. Real-time estimation of freeway accident likelihood[C/CD]//80th Annual Meeting of the Transportation Research Board. Washington D C: Transportation Research Board, 2001.

[2] Lee C, Hellinga B, Saccomanno F. Real-time crash prediction model for application to crash prevention in freeway traffic[J]. Journal of the Transportation Research Board, 2003, 1840(1): 67.

[3] Abdel-Aty M, Uddin N, Pande A, *et al*. Predicting freeway crashes from loop detector data by matched case-control logistic regression[J]. Journal of the Transportation Research Board, 2004, 1897(1): 88.

[4] Abdel-Aty M, Abdalla F M. Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes: generalized estimating equations for correlated data [J]. Journal of the Transportation Research Board, 2004, 1897(1): 106.

[5] Ahmed M M, Abdel-Aty M A. The viability of using automatic vehicle identification data for real-time crash prediction [J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(2): 459.

[6] Abdel-Aty M, Pande A. Identifying crash propensity using specific traffic speed conditions [J]. Journal of Safety Research, 2005, 36(1): 97.

[7] Abdel-Aty M, Pande A. Classification of real-time traffic speed patterns to predict crashes on freeways[C/CD]//83rd Annual Meeting of the Transportation Research Board. Washington D C: Transportation Research Board, 2004.

(下转第 324 页)