文章编号:1671-4598(2016)04-0230-04 DOI:10.16526/j.cnki.11—4762/tp.2016.04.066 中图分类号:TP39 文献标识码:A

集成随机森林的交通拥堵检测模型

谭 娟[1], 王胜春[2]
(1. 北京工商大学 商学院, 北京 100048; 2. 北京交通大学 交通数据分析与挖掘北京市重点实验室, 北京 100044)

摘要: 根据现有的城市交通网拥堵检测体系, 针对现有方法处理交通网格监测数据流难以获得相对稳定的准确率的问题, 提出了一种集成随机森林的交通拥堵检测模型; 该模型通过将多个随机森林分类器进行集成实现了交通网分布式监测数据流的并行处理, 设计了二级级联分类器对交通网状态进行判定, 并可对各监控节点权重进行评估; 模型实现主要分为特征提取、集成建模和结合分析3个步骤; 在不同规模的交通状态监测网络下分析了模型的综合性能, 并分别与其它主流方法进行了对比; 实验表明: 提出模型具有更好的交通网监测数据流的处理能力, 且具备较好的扩展和裁剪性能; 该模型提供了一种可应用的交通拥堵检测方法。

关键词: 交通拥堵检测; 随机森林; 级联分类器; 节点权重

Traffic Congestion Detection Model Based on Ensemble Random Forests

Tan Juan[1], Wang Shengchun[2]
(1. Business School, Beijing Technology and Business University, Beijing 100048, China;
2. Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China)

Abstract: According to the existing traffic congestion detection system of cities, a detection system is proposed to solve the problems of relatively low and unstable accuracy in processing the traffic monitoring data. This model integrated multiple random forests (RF) to process each node data in the traffic network parallel, then a cascade classifier is designed to recognize the traffic network status. At last, the importance of node in the traffic network is assessed by using RF. The implementation of this model mainly consisted of three levels, that is, feature extraction, building the integrated classification model and combination analysis. Comprehensive performance of the model is analyzed under different size traffic network. and compared respectively with other algorithms. Finally, experiments show the proposed model not only has better comprehensive performance in traffic network monitoring data, but also can be adapt to the change of network size. This model provides an application model for traffic congestion detection.

Keywords: traffic congestion detection; random forest; cascade classifier; weight factor node

0 引言

交通拥挤是世界各大中城市所面临的共同问题。由于汽车保有量逐年持续增加, 以及交通信息供给不足和管理措施不利, 造成了现有的大中城市在行车高峰期的交通效率低下, 给整个社会发展带来了一系列经济、安全、环境污染等多方面的问题。解决交通拥挤的传统办法是拓宽道路或建设新路。但受限于城市土地面积及规划, 采用这种方法解决交通拥挤的难度越来越大。另外分配道路通行能力虽然可以缓解交通拥挤, 但是它存在着经济效率差、影响出行者时间上的公平性等缺点。因此, 构建交通拥堵检测及交通诱导信息系统成为缓解城市交通拥堵的有效手段。

交通拥挤的检测及交通诱导信息的发布是依靠传感器监测、信息处理技术、通信技术来共同实现。目前我国大中城市(以北上广深为例)交通疏导监测体系的特点是传感器布设密集, 但交通信息处理技术相对单一, 多采用单点分析、统计分析结合带阈值的多数投票法[1]进行交通状态判别。这些方法存在对于网格化监控数据处理的集成化程度低, 对交通网内造成拥堵节点的综合评判正确率不高, 以及数据有效信息挖掘深度不够的问题。这已成为这类系统发挥正常功能的瓶颈, 易导致事件检测率偏低等问题。

目前, 有许多学者结合大数据思维, 开展城际高速公路、城市公路交通等路网监测数据进行综合集成处理的研究, 以期获得对较好预测交通拥挤状态的检测模型。不同的拥堵检测模型尽管在拥堵判别的准则上存在差异, 但其理念都是依据交通流监测数据的建模分析实现。文献[2-4]就神经网络算法在交通拥堵预测中的应用进行了研究, 但神经网络方法的参数选取缺少依据, 且对训练过程有较强依赖, 模型识别性能不稳定。文献[5]提出了基于ID3方法构建的决策树分类, 但该方法受限于数据分布, 容易出现过拟合的问题。近年来, 随着数据挖掘技术发展, 部分学者把支持向量机(SVM)方法应用于构建交通拥堵监测模型[6-7], 能获得90.6%的判别精度, 较好的推动了交通监测数据分析技术的发展。在交通拥堵分析模型的研究方面, 部分学者选择从其它更多的视角[8]和分析方法力求获得对交通瞬时数据更高的处理精度, 比如: 基于MapReduce方法[9], 基于影响模型[10]方法等。

不同地, 本文提出了一种集成随机森林的交通拥堵检测模型。随机森林(random forests, RF)是由Leo Breiman提出, 并由诸多数学家在后续研究中不断完善的模式分类算法。它实际上是一个由一系列决策树形式的基础分类器以随机方式构建的组合分类器[11]。该方法适合于处理具有高度相关特征的高

收稿日期:2015-10-27; **修回日期**:2015-11-19。
基金项目:北京市自然科学青年基金(9144022);北京市社会科学基金项目(15JGC159);首都流通业研究基地支助项目(JD—YB—2016—004)。
作者简介:谭 娟(1983-),女,湖南邵阳人,博士,副教授,主要从事交通运输规划、环境经济管理方向的研究。

维数据集，其变量重要性度量为特征选择提供了一个自然方法。本文基于随机森林作为基础算法开展研究的优势在于[12]：1）决策分类的理论支撑更为完善，具有更好的分类精度，且不会产生过拟合问题；2）能够有效地处理大数据集和高维数据集，适用于交通路网监测数据分析；3）能够在分类过程中对特征变量对分类决策影响的重要性进行估计。借助该项能力，可实现对路网各节点对交通网总体状态的影响程度评估。

本文根据交通监控数据的网格分布属性，构建了一个集成随机森林的交通拥堵检测模型。首先采用随机森林对各监测节点数据进行处理，并获得节点状态的初次分类。然后将各节点的状态输出结果进行二次组合，再次采用随机森林算法构建一个总分类器对路网状态进行判定，并给出各节点权重的影响评估方法。论文将在第 2 部分阐述集成随机森林的设计和建模过程。第 3 部分阐述测试情境，训练集和测试集的构造，重点展示数据并分析结果，并给出节点权重分析的影响方法。

1　集成随机森林建模

1.1　监控数据采集和特征建模

所谓交通拥挤，是指一定时间内道路的交通需求超过其通行能力或者由于突发交通事件造成道路通行能力短时下降并低于当时的交通需求而发生的交通流滞留在道路上的交通现象。对交通状态的定义按照以下三类进行定义：

1）通畅（tc）：道路的需求量低，路上车辆较少，出行者可以很快到达目的地；2）正常（zc）：由自由流状态逐渐转化为间歇性停滞，但通行状态基本可控；3）拥堵（yd）：车辆走行缓慢，来自上游的交通需求中无法通过瓶颈，形成排队。本文将交通拥挤定义为 5 min 车道占有率超过 30%的交通状态。

交通拥堵状态分析数据的获取需以现有的交通监测体系为依据。交通监测最常采用的分析模型是网型结构，如图 1 所示。在交叉路口的 4 个通行方向均设立测点，在部分主干道长直线路段的中间也设置测点。本文以各节点的监测数据为研究对象，该网络中各节点之间的数据具有直接（如 1、2、3、7 节点）或间接（如 1、5 节点）相关性。

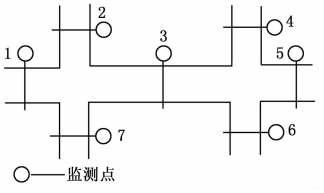

图 1　交通监测体系的网型结构

依据常理，在路网中通行的车流方向变化规律是难以预知的。因此要根据当前监测数据检测路网交通状态的变化情况，通过建立数据挖掘模型研究历史数据的潜在规律是解决问题的手段之一。本文根据随机森林方法的若干优势，以该方法为基础，建立集成模型开展研究。

根据布设的传感器并进行统计分析建模，对数据进行合成处理后定义 4 个参数作为基础变量：1）流量（fq）：指单位时间（1 分钟）行驶通过道路指定地点监测断面的车辆数；2）空间占有率（spr）：在道路的一定路段上，车辆总长度与路段总长度之比称为空间占有率；3）车头时距离（td）：对前后两车通过车行道上某一点的时间差的测量值；4）时间平均速度（mv）：是指在特定的时间区间内，通过道路某一地点的所有车辆点速度的算术平均值。

与传感器监测参数对应的物理变量相对应，按式（1）定义每个节点的输入、输出特征向量：

$$\begin{aligned}&\mathrm{In}:X_{input}=[fq,spr,td,mv]\\&\mathrm{Out}:Y_{output}=[tc,zc,yd]\end{aligned}\tag{1}$$

对于集成模型，模型的原始输入由对应不同节点的 N 个观测组 $L=\{(x_i,y_i),i=1,2,\cdots,N\}$，每组样本对应一组如式（1）所示向量。

1.2　集成模型

本文通过建立集成模型分析并处理交通路网监测数据。模型设计从两个角度切入：1）期望能较好地并行处理各单点监测数据，可根据点数据流判断单点的交通状态；2）能对各节点组成的路网的综合状态进行全局判定，并在一定程度量化给出网内各节点状态的影响评价。

集成模型的构建方式分三步实现。其总体结构如图 2 所示。模型设计采用二级级联结构。第一级为多个单点分类器的并行结构处理。第二级模型以第一级的判定输出标签作为输入集，构建对路网综合状态的判定模型。模型的核心算法基于随机森林实现。

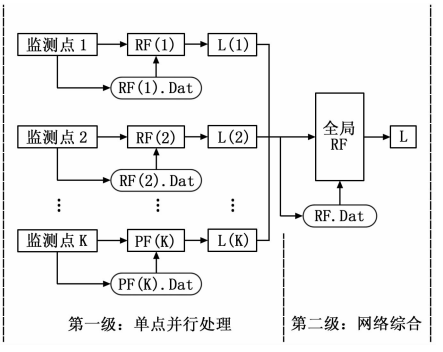

图 2　集成模型总体框架图

第 1 步：数据集构造和输入。该步骤包含数据预处理和分类过程。数据预处理过程需多次采集训练和测试所需的数据，然后计算相关特征，构成特征集。在数据挖掘模型建立之前，对数据做归一化真值处理。根据经验法设定阈值，选定参数建立分段阈值函数（如式（2）所示），将测量值进行分级标注，转化为特征真值表作为训练集。

$$X_{\mathrm{var}}=\begin{cases}V_1 & x<th1\\ f(x) & th1\leqslant x<th2\\ V_2 & x\geqslant th2\end{cases}\tag{2}$$

第 2 步：训练和建模。该步骤分两个阶段实现。在步骤 1 的基础上，首先建立第一级模型。每个模型是一个独立的随机森林模型。训练过程如下：

1）给定训练集 L，测试集 T，特征维数 $M=4$。设定参数：确定模型中决策树的数量 $ntree$，树的深度 d，每颗树节点使用到的特征数量 $mtry$；并设定训练的终止条件：树节点上样本数最小值 s，节点上信息增益最小值 f；

2）从 L 中有放回的提取新训练集 L（i），保证 L（i）的样本规模和 L 一致，并作为根节点的样本集开始训练。

3）判断当前节点是否达到终止条件。如是，则设置当前节点为叶子节点，然后继续训练其他节点；如否，则从 M 维特征中无放回地随机选取 f 维特征。利用这 f 维特征，寻找

分类效果最佳的特征 m 及其阈值 th：在当前节点，样本第 m 维特征小于 th 的样本被划分到左节点，否则划分到右节点。然后继续训练其他节点。

4）重复 2）、3）直到遍历节点，或被标记为叶子节点。

5）重复 2）～4）直到遍历所有决策树。

随机森林的决策树分类效果评价采用 Gini 标准值，其计算公式如式（3）：

$$\begin{cases} Gini = 1 - \sum P(i) * P(i) \\ \text{评判标准：} \underset{f,th}{\operatorname{argmax}} \ (Gini - Gini_L - Gini_R) \end{cases} \tag{3}$$

式中，$P(i)$ 为当前节点上数据集中第 i 类样本的比例。评判标准为：寻找最佳的特征 f 和阈值 th，使得当前节点的 $Gini$ 值减去左右分支节点的 $Gini$ 值最大。

在模型的整体构建上，对应 K 个监测节点分别建立 K 个基于随机森林的模型，分别为 RF（1）～RF（K）；并将每个随机森林模型的训练参数和模型文件 RF（1）.dat～RF（K）.dat 存储到磁盘。

第二级模型的建立以第一级模型所有节点的输出作为输入，建立新的随机森林模型分析各节点交通状态所形成的综合效应。训练过程与第一级模型的训练过程类似，但模型特征维数 M 的设定根据节点数量 K 进行设定。模型的训练集和测试集的建立采用第 1 步中所述数据集构造方式获得。对于路网交通状态监测节点的评价依据经验确定。该步骤训练完成后，获得一个随机森林模型文件 RF. dat。

在预测分类时，各监测节点之构成分布式计算网络。模型通过实时加载预测数据和各步骤已生成的模型文件，通过随机森林决策树输出交通状态。在实际处理时，各节点在某一时刻 t 同时获得监测数据，经模型处理后输出获得 t 时刻的交通网络状态标签。

第 3 步：综合分析。该步骤是指根据数据处理模型的连续输出状态，设计规则对路网状态进行最终评判并预警。由于路网通行状态是一个动态变化的过程，因此需根据模型预测结果观测值的连续变化情况来判定路网的状态。较为常用的设计方法是：定义以时间为自变量的状态统计函数，并以量化值作为输出分级描述交通拥堵状态（即拥堵指数）。该函数的定义更为依赖模型的应用情境，在此不对具体形式进行定义。

同时，利用随机森林分类效果评价标准所提供的特征重要性度量方法，本文设计的第二级模型可通过对数据集的处理，对网络节点的影响因数进行量化评估。从数据挖掘角度阐述，影响因子越高，说明该变量对于随机森林分类器获得更高正确率的贡献度越高。对于交通网数据分析而言，说明该节点对整个交通路网中的影响力越高。根据模型的向量评价系数进行排序，可以对网络全部节点的影响程度进行观测，以获得对交通网络进行疏导改造的建议。

1.3 参数调试方法

随机森林模型的参数较多，当模型性能不满足要求时，需要对模型的参数进行调整，使模型的误差率减小，达到更好的性能。其中最主要的两个参数为：*ntree*（随机森林中树的数目）、*mtry*（节点随机选择的特征数目）。

（1）参数 *ntree* 的确定方法。随机森林基础理论证明，只要 *ntree* 足够大，模型误差将达到一个固定上限；但树的颗树如果过多，又会损失效率。该参数调试需结合实际数据集进行动态观测调试，确定系统到达误差稳态上限时的临界区。

（2）参数 *mtry* 的确定方法。参数 *mtry* 是指随机森林模型中决策树除了根节点、叶节点以外的其他节点处随机选择特征的数目，*mtry* 比其他参数对模型的性能更加敏感。本文对该参数的计算采用式（4）：

$$mtry = [sqrt(M)] \tag{4}$$

即该参数的取值为输入向量特征数 M 开平方并向下取整。对于本文模型，第一级中各测点的 RF 模型取值为 2。第二级网络综合 RF 的 *mtry* 参数取值根据网络节点数量确定。

1.4 评估指标

随机森林模型可在分类器训练过程中自动生成“袋外”（Out-of-bag）样本集用于估计模型的泛化误差。因此本文对 RF（n）、RF 模型的性能评估均以模型输出的 out-of-bag 样本集估计误差率（*oobE*）为依据[13]。因此对于一级模型的单节点随机森林模型而言，该误差率参数取值越小，表示模型性能越好。对于模型的总体精度而言，通过考察模型对测试集处理的准确率来评估性能，准确率定义如下：

$$\text{模型分类准确率} = \frac{\text{分类正确的样本数}}{\text{样本总数}} \times 100\% \tag{5}$$

2 模型验证及分析

本节从 3 个方面设计对试验进行验证。主要试验工作如下：1）通过观测 RF 模型中 ntree 参数与 oobE 误差率的变化关系确定 ntee 最佳取值；2）通过建立不同规模的网络，验证模型对交通路网监测数据的处理精度和处理能力；3）将本文模型与 BP 神经网络（BP-NN）、线性支持向量机（LSVM）进行比较，分析模型在交通监测处理能力上的差异。

2.1 数据采集与整理

本研究采用北京市朝阳区 CBD 区域为中心的路网监测数据构建数据集进行测试验证。数据采集和整理方法如下：（1）依据该区域交通在不同时间段出现的整体拥堵状态，按照定义的 3 种交通状态进行分类整理；（2）与本文提出的二级模型相对应，对每个监测点在不同时间段的交通状态变化状况也进行统计；（3）数据搜集考虑了周一早高峰、周五晚高峰等可能出现的极端拥堵情况，但不考虑交通状态极度通畅的情况，因此对凌晨 0 点到 6 点的数据予以剔除。

2.2 试验环境

测试数据集的整理、真值化处理，模型的训练模块和分析处理模块的编程实现均在 Matlab（版本：2012b）平台下完成。工作 PC 机的操作系统为 Windows7（64 bit），硬件配置为 CPU：Intel-I7，内存：32 G，硬盘：3 T。

2.3 结果和分析

2.3.1 *ntree* 参数与 *oobE* 误差率的变化关系

该项验证设定随机森林模型训练函数 *ntree* 参数的上限为 50，并通过观测模型在 *ntree* 参数变化时 *oobE* 误差率的变化值。结果如图 3 所示。这里分别考察 2 个一级模型中单测点 RF 模型和网络综合 RF 模型的变化状态，从图上可以看出，*ntree* 参数取值在［2，10］区间时，误差率随 *ntree* 参数增大出现了急剧衰减；在［10，20］区间，则呈现平缓衰减状态，并在之后逐渐达到稳态。系统稳态误差率 *oobE* 的平均值小于 3%。综上所述，本文一级模型各测点 RF 模型的 *ntree* 参数的理想值区间为［20，25］，二级模型 RF 模型的 *ntree* 参数的理想区间值在［25，30］。

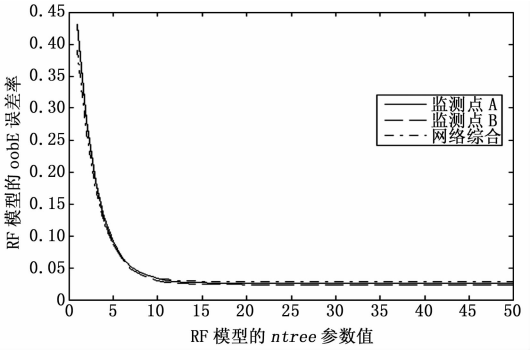

图 3　*ntree* 参数与 *oobE* 变化关系曲线图

2.3.2　模型的综合性能分析

以验证模型的综合处理能力为目标，根据不同监测点的数量进行组网，节点数目的规模分别为 14、20、26、32。根据不同的网络规模分别训练模型，然后以测试数据集作为输入，并校验模型输出的准确率。为了便于比较算法性能，在相同的数据集上同时采用了 BP 神经网络（BP-NN）和线性支持向量机（LSVM）进行处理，与本文模型的对比结果如图 4 所示。

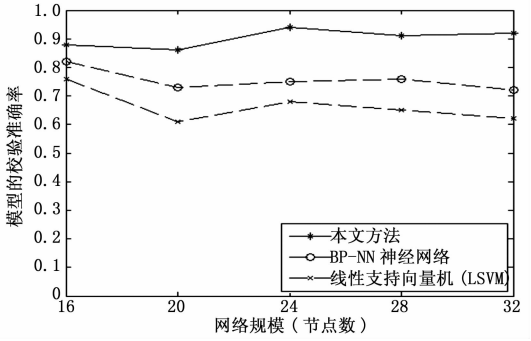

图 4　本文方法与其他方法的数据处理性能对比

从对比结果可以看出，本文方法的平均准确率高出 BP-NN 方法 15 个百分点左右，而 LSVM 则表现的不太理想。在处理不同规模的网络时，随着网络节点数量的增加，本文方法表现出的性能较为稳定，而其它两种方法则在精度上开始出现衰减。从设计机理分析，由于本文方法对各监测节点采用了分级方法进行处理，使得系统在适应网络和数据规模扩展方面具备了更为优良的性能。

2.3.3　各监测点影响权重评估分析

随机森林算法的另一个特点是能在分类过程中对模型变量的影响权重进行评估。本研究利用算法的该项能力，在第二级处理的 RF 模型中对交通路网中各监测节点对获得高分类正确率的影响程度进行计算。该值从一定程度上可以反应各监测点位置在整个交通路网中的影响程度。这里以 B 部分中 32 节点的交通网络为例进行分析，并给出排名前 5 的测点情况。具体数据见表 1。

二级模型中的 RF 方法对每个监测点都给出了一个数值评估，该值是一个对模型输入变量的评估系数，通过排序分析可观测其实际参考意义。将排序结果与监测点所在位置进行对照分析，可以看出排名靠前的监测位置表现出了一定程度的聚集性。对应路网的实际情况分析，这些位置可以视为交通通行的

表 1　各监测点在路网中的重要性评估分析

序次	测点	重要性权重	监测位置
1	测点 07	8.578 5	a 路口－左
2	测点 15	8.266 7	b 路口－右
3	测点 08	7.428 5	a 路口－上
4	测点 12	7.156 9	b 路口－上
3	测点 13	6.684 1	b 路口－下

“瓶颈”位置。由此可见，本文方法为交通网络节点状态的综合评估提供了一种量化分析体系。

3　结束语

交通网络拥堵状态的检测模型是城市智能交通技术研究的热点问题。本文提出的集成随机深林的处理模型采用二级级联结构设计。第一级模型采用 RF 方法对交通路网监测数据进行并行集成处理，第二级模型依据各监测节点状态建立 RF 分类器，实现了对交通路网状态的综合分析。经试验论证，本文模型在处理交通路网监测数据时的准确率高，同时兼备了对单点数据和网络数据进行分析挖掘的综合性能力，是一种可以应用于构建交通拥堵检测及诱导信息系统的数据挖掘模型。同时，该建模方法可供其它应用背景的网格数据挖掘分析方法借鉴。

参考文献：

[1] Xiang G Y，Niu S F，An D C. The method of traffic congestion identification and spatial and temporal dispersion range estimation [A]. International Asia Conference on Informatics in Control，Automation and Robotics [C]. 2010，1：36－39.

[2] 陈阳舟，田秋芳，张利国. 基于神经网络的城市快速路交通拥堵判别算法 [J]. 计算机测量与控制，2011，19 (1)：167－169

[3] 杨祖元，黄席樾，杜长海，等. 基于 FFCM 聚类的城市交通拥堵判别研究 [J]. 计算机应用研究，2008，25 (9)：2768－2770.

[4] 鲁小丫，宋志豪，徐　柱，等. 利用实时路况数据聚类方法检测城市交通拥堵点 [J]. 地球信息科学学报，2012，14 (6)：775－779.

[5] Srinivasan D，Sanyal S，Sharma V. Freeway incident detection using hybrid neural network [J]. IEEE Transaction on Intelligent Transportation System，2007，1 (4)：249－259.

[6] 郭　倩，黄　林. 基于粗糙集和支持向量机的高速公路事件检测 [J]. 计算机工程与应用. 2008，44 (35)：203－205.

[7] 郑长江，路　源. 基于支持向量机的城市道路交通拥堵判别算法研究 [J]. 贵州大学学报（自然科学版），2014，31 (1)：113－117.

[8] 杨聚芬，姜桂艳，李　琦. 基于收费数据的高速公路交通拥挤自动判别方法 [J]. 哈尔滨工业大学学报，2014，46 (12)：108－113

[9] 梁　轲，谭建军，李英远. 一种基于 MapReduce 的短时交通流预测方法 [J]. 计算机工程，2015，41 (1)：174－179.

[10] 丁　栋，朱云龙，库　涛，等. 基于影响模型的短时交通流预测方法 [J]. 计算机工程. 2012，38 (10)：164－167.

[11] 华　楠. 基于数据挖掘技术的交通拥挤检测及应用 [D]. 长春：吉林大学，2008.

[12] 邓生雄，雒江涛，刘　勇，等. 集成随机森林的分类模型 [J]. 计算机应用研究，2015，32 (6)：1621－1625.

[13] 张春霞，郭　高. Out _ of _ bag 样本的应用研究 [J]. 软件，2011，32 (3)：1－4.