

Forecast of Short-term Traffic Flow in Chengdu City Based on Decision Tree Model

Jingdian Yang

Chengdu College of UESTC, 610097, China

Email: 492869063@qq.com

Abstract

With the rapid increase in the population in the central area of Chengdu, congestion in a certain period has become one of the stubborn diseases that restrict the development of the city. By accurately predicting the short-term road traffic information in the future, it can effectively improve the traffic efficiency of vehicles in a specific period. In this paper, data training is carried out according to the collected relevant indicators, and a decision tree classification model is established. The model can be used to predict the traffic flow on the road in a short period of time based on the time of the road, weather, average speed, and other indicators, and then the relevant regulatory authorities can use the advanced new media today to push, so that the citizens can plan their trips reasonably.

Keywords: Traffic Jam; Machine Learning; Decision Tree; Prediction

基于决策树模型的成都市短时交通流预测

杨京典

电子科技大学成都学院, 四川省 成都市 610000

摘要: 随着成都市中心区域人口的快速增加, 在特定时段的拥堵现象已经成为限制城市发展的顽疾之一, 通过准确的预测未来短时路面的交通信息, 能有效的提高特定时段的车的通行效率。本文中根据采集到的相关指标进行数据训练, 建立决策树分类模型。使用该模型可根据道路的时段、天气、平均速度等指标预测出短时内该路面的交通流情况, 随后相关监管部门可以利用当今社会发达的新媒体进行推送, 让市民合理规划出行。

关键词: 中文交通拥堵; 机器学习; 决策树; 预测

1 引言

交通拥堵状态是用来评价城市路面行驶效率的指标。美国国家公路交通安全局(NHTSA)^[1]曾发布《道路通行能力守则》。该守则将路面的拥堵状态分为了六个级别, 包括: 非常畅通、畅通、轻度拥堵、严重拥堵、锁死。而我国北京交通发展研究中的相关报告^[2]也指出当前北京市的道路交通状态可以分为五个等级, 并建立了模型分析对应等级对居民出行时间的影响。结果表明, 当路面的车流等级为严重拥堵时, 居民单次出行所花的时间为畅通状态的一倍。因此拥堵的交通会极大的降低城市运行效率, 为城市的发展带来损失。

因此, 如何缓解交通拥堵, 已成为时下最热门的研究之一。在大数据时代, 数据挖掘学作为当今人工智能中的热门学科, 其中的一项任务预测受到了研究人员的广泛关注。预测任务的目标是根据其他属性的值, 预测特定属性的值。被预测的属性一般称目标变量(target variable)或因变量(dependent variable), 而用来做预测的属性称为说明变量(explanatory variable)或自变量(independent variable)。

分类是预测任务中的一种有监督式学习，可以看成是从一个数据集到一组预先定义的、非交叠的类别的映射过程。即从给定的有标记训练数据集中学习出一个函数，当未标记数据到来时，可以根据这个函数预测结果。常见的分类模型包括决策树、数学公式、神经网络等。本研究选择了决策树模型，根据已有的数据对未来短时道路交通的拥堵情况做出预测，再利用当今社会发达的通讯方式如短信、微信微博公众号通知出行居民，建议其选择公共交通或者道路，能够很好的提升城市交通的运转效率。

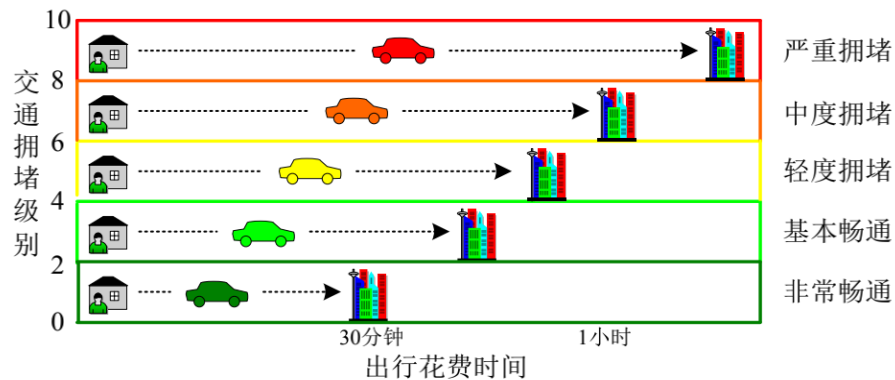


图1 出行时间与交通拥堵的关系

2 拥堵指标选择与模型建立

2.1 影响道路交通指标选择

针对交通拥堵展开的相关预测已有很多，晏雨婵^[3]针对各交通时段对交通拥堵的不同影响、单因素无法准确表征交通拥堵状态的问题，提出了一种采用多指标模糊综合评价的交通拥堵评价预测方法，具有较高的预测准度。刘张^[4]等人为了更高的预测精确度提出了一种添加调节项的高阶多变量马尔可夫模型，证明了模型的收敛性，进行了参数估计，并参考城市道路交通运行评价指标体系，对城市拥堵进行预测分析。支野^[5]等人探索了道路交通数据缺失数据项填补、事故伤亡特征因子甄别以及事故危险性分类评价方法。结果表明，该方法可有效提高道路交通 事故数据完整性和事故伤亡特征因子甄别准确性，以及量化交通事故危险度评价。在本研究中，数据采集时考虑了五个不同的因素，包括了时间段、天气、路况、节假日和特殊事件。

2.2 决策树 ID3 和 CART 算法

分类的方法在当今已经非常成熟，如常见的朴素贝叶斯、SMV、KNN 以及决策树等等，这些方法在解决问题的能力上各有所长。如何利用这些模型实现更高的预测准确度便是目前研究的重点。其中决策树模型从提出到现在一直都是数据挖掘中的热门模型，它代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值。所以它不需要了解很多的背景知识，非常便于使用者理解和实现。常见的决策树^[6]算法包括了 ID3、C45 和 CART 算法，其中 ID3 算法是目前决策树中历史最悠久，也最成熟的一种，它的核心是在决策树各级节点上选择属性时，用信息增益作为属性的选择标准，使得在每一个非节点进行测试时，能获得关于被测试记录最大的类别信息。

要了解信息增益这一概念则需要先知道信息熵，熵的概念比较复杂，它来源于热力学，代表世间万物的混乱程度。1948 年香农将熵的概念引入到了信道通信中，从而开创了信息学这门新的学科，信息熵的概念实现了对于信息的量化，其数学概念如下：

假定当前的样本集合 D 中第 k 类样本所占的比例是 P_k ；

那么 D 的信息熵定义为：

$$\text{Entropy (D)} = - \sum_{k=1}^{|y|} P_k \log_2 P_k$$

由上面可知, $p=0$, 则 $P_k \log_2 P_k = 0$, $\rightarrow$ 也就是说 Entropy (D) 的最小值为 0, 最大值是 $\log_2 P_k$ 。且 Entropy (D) 的值越小, 则 D 的纯度就越高。

信息增益直接基于信息熵, 计算当前划分引起的信息熵变化, 说明选择特征 A 降低了数据集分类的不确定性, 越降低, 数据集分类的不确定性越低。特征 A 对数据集 D 的分类效果越好, 信息增益针对的是一个特征, 也就是看一个特征, 系统有和没有的时候信息量是多少, 两者的区别是这个特征给系统带来的信息量, 也就是信息增益。在决策树问题中, 信息增益是属性选择和划分前后的信息之差。一个节点的信息增益越大, 使用属性 A 进行划分得到的“纯度提升”越大。在划分决策树的每个非叶节点之前, 首先计算每个属性带来的信息增益, 选择信息增益最大的属性进行划分, 因为信息增益越大, 区分样本的能力越强, 越有代表性。

决策树 CART 算法使用了基尼系数来代替信息增益, 基尼系数代表了模型的不纯度, 基尼系数越小的时候, 模型的纯净度越高, 这和 ID3 算法的信息增益度量方式相反基尼系数和熵之半的曲线非常接近, 因此, 基尼系数可以做为熵模型的一个近似替代。在本研究中, 会使用 ID3 和 CART 算法的模型做对比测试, 验证两种算法的预测准确率。

2.3 拥堵标准评判

在道路交通状况的评判中, 常见的分类即拥堵和通畅两种情况, 但如何划定拥堵与通畅的界限则需要具体的数据作为量化的指标, 本研究根据国外的相关研究结果, 可使用一个交通指数的变量 α 来表示道路的拥挤情况:

$$\alpha = \text{交通量} / 300$$

其中 α 的取值越大, 就表示在路上耽误的时长更多, 所以道路就越拥挤。其具体的规定如下:

$\alpha \in [0, 1)$ 时道路状态为“畅通”,

$\alpha \in [1, +\infty)$ 时道路状态为“拥堵”。

通过计算 α 的值, 我们可以对采集到的数据完成分类, 判定其为拥堵或者通畅。

3 实验例证与性能评估

3.1 数据采集与预处理

本研究中使用的数据为人工采集, 地点位于成都市高新西区, 路段名称为百草路、新航路和新业路, 数据采集的时间周期为 2021 年 1 月 1 日至 2021 年 1 月 16 日共计 16 天, 采样周期间隔为 30min, 每个路段每天的数据条数为 25 条, 一条路 16 天共采集到 400 个数据。

将含有缺失数据、错误数据的那一条数据直接删除, 然后用缺失的数据相邻两条求出平均值填补那条错误的数, 即采用相邻数据平均法补全, 若对于明显错误的数, 比如单位时间内的流量为个位数或负数^[7], 则按照缺失数据处理。在数据预处理完成之后, 利用可视化软件得到以下流量图:

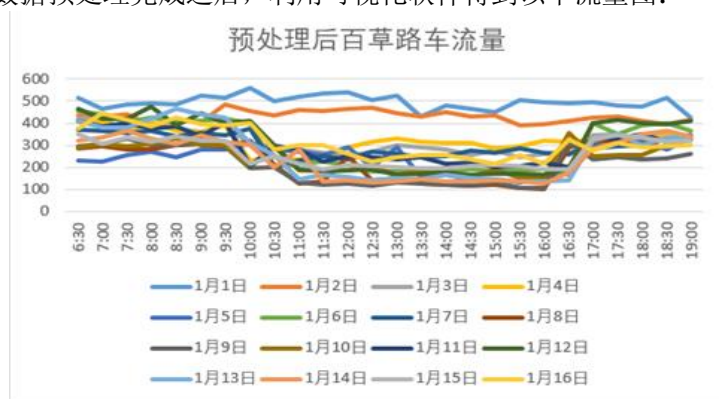


图 2 百草路短时交通流量图

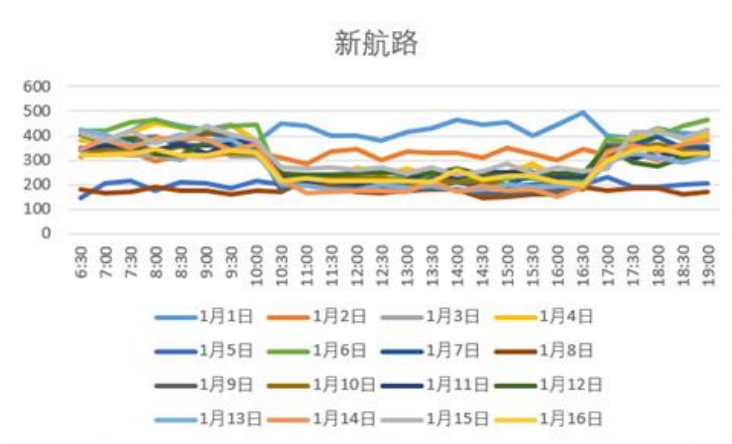


图 3 新航路短时交通流量图

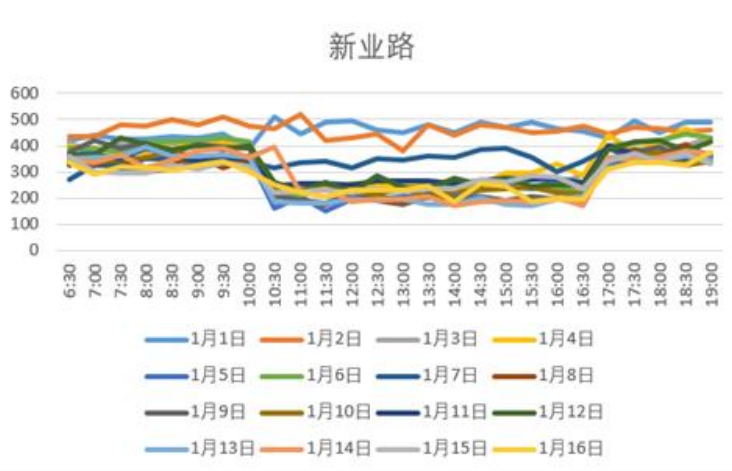


图 4 新业路短时交通流量图

从可视化的流量图像中我们也可以看出：两条不同的道路在周末两天和平时的交通流量上面趋势大致相同，早高峰的时间段集中于 6:30-10:00，晚高峰的时间段在 17:00-19:00，并且在周末的时候各个时间段出行量并无太大差别，这种规律性为短时交通流预测提供了良好的预测保障。

3.2 数据挖掘和算法实现

本研究使用的是 Python 语言完成的数据挖掘工作，使用的平台为 pycharm 集成开发工具，利用 Python 第三方库 sklearn 中强大的功能模块完成的预测与模型性能测试工作，其具体的实现工作步骤为：

1. 利用 `train_test_split` 方法将数据分割为训练集与测试集，按照一比一的比例进行划分；
2. 利用 `DictVectorizer` 类进行特征的提取，在本次研究中提取了其中三个特征；
3. 使用 `DecisionTreeClassifier` 算法模块对数据进行分类；
4. 利用 `export_graphviz` 进行决策树的绘制，并转换为 PNG。

在实现以上步骤的过程中，首先选取了三个特定字段-天气、节假日和特殊事件，在决策树的构造中设置最大深度为 7，在 `DecisionTreeClassifier()` 中设置算法类型为 `entropy` 即使用决策树 ID3 算法。另外，为了使得最后生成的决策树各个属性更加清晰，因此将部分属性值进行了缩写。最终将运行得到模型准确率以及生成一个决策树模型图的 `dot` 文件，将文件转换成图片即可看到如下一颗决策树：

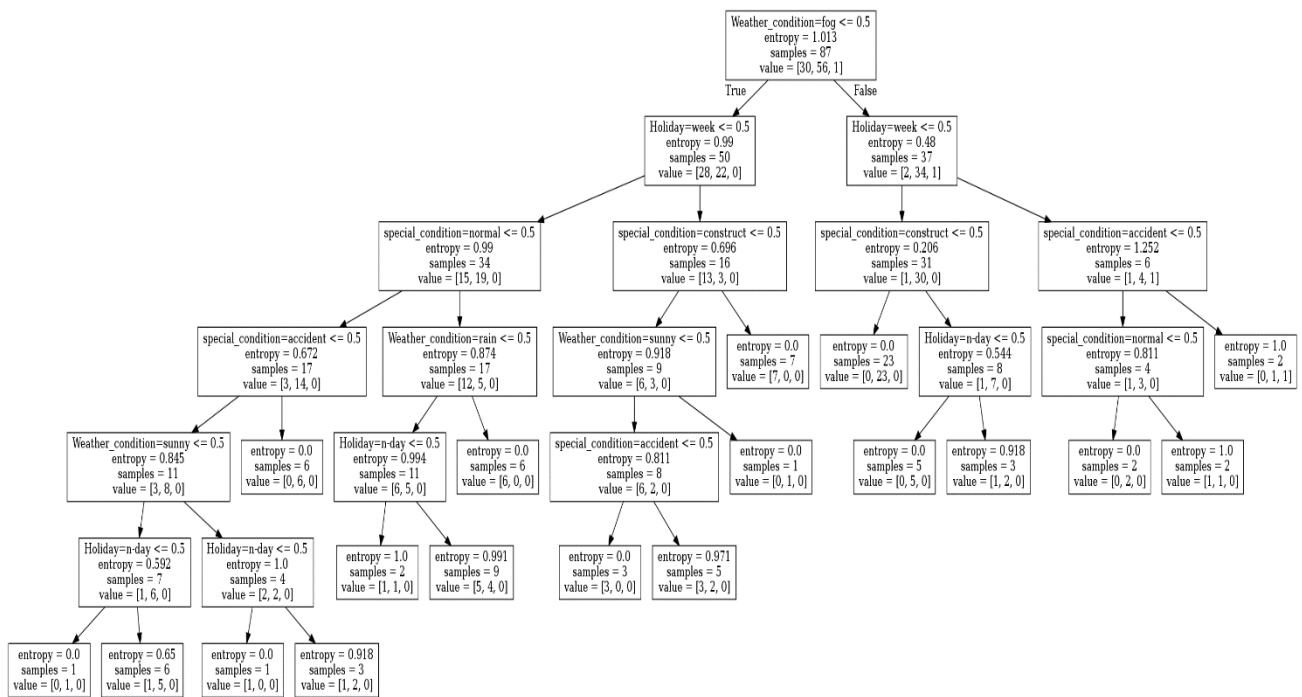


图 5 使用 ID3 算法运行生成的决策树和预测准确率

将算法的模式改为 gini，同样设置树的最大深度为 7，根据 gini 系数分类决策树生成了 31 个节点对应 30 条规则路，然后同样通过 10 次运行随机划分数据的测试集和训练集，模型的准确率出现了较大的浮动，最低为 61%，最高为 77%，平均为 70%，表明其预测性能不如 ID3 算法。

```
Index(['Time', 'Weather_condition', 'Road_c',
      'special_condition', 'Result'],
      dtype='object')
['Holiday=n-day', 'Holiday=non-week', 'Holi
```

决策树模型预测的准确率为: 0.7159090909090909

图 6 使用 ID3 算法运行生成的决策树和预测准确率

4 结论

对比高德地图的实时路况播报,在本研究中,读者提出了利用决策树的 ID3 和 CART 算法对采集的数据进行了训练与模型建立,并做出了最终的短时交通流拥堵预测。相比实时的路况播报具有更高的先进性与实用性,本研究在经过人工采集了交通流数据城市道路的交通流数据之后,可以分离出交通状态的影响因素,即其不同状态不同天气不同道路情况下的拥挤状态,通过离散化和标准化的处理之后,将其用于决策树分析,建立了成都市短时交通流预测模型和规则,用于预测出行道路的近期拥挤状况,从实验的结果可以看到, ID3 算法的性能明显更好。同时交管部门和将要出行的人可以根据本次实验的预测结果采取相应的措施以应对道路不同情况对于自己所产生的负面影响,在引入了百草路、新业路、新航路三条主干道的数据之后,实验表明具有良好的预测准确率,具有较高的实用价值。

参考文献

- [1] 易斌,宋程,刘翰宁.基于大数据的交通拥堵影响因素关联性实证分析[J].交通与运输,2021,37(02):30-35.
- [2] 郑苗,吕永艺.城市道路中智能交通应用研究[J].智能建筑与智慧城市,2021(03):130-131+137.
- [3] 晏雨婵,白璘,武奇生,叶珍.基于多指标模糊综合评价的交通拥堵预测与评估[J].计算机应用研究, 2019,36(12)
- [4] 刘张,李坚,王超,蔡世民,唐明,黄琦,陈照辉. 基于复杂城市道路网络的交通拥堵预测模型[J]. 电子科技大学学报. 2016,45(01).
- [5] 支野,王大珊,丛浩哲,饶众博. 道路交通事故数据深度挖掘技术与应用——以深圳市为例[J]. 城市交通. 2018,16(03).
- [6] 薛红军,陈广交,李鑫民,顾理.基于决策树理论的交通流参数短时预测[J].交通信息与安全,2016,34(03).
- [7] 于田. 基于机器学习的交通流缺失数据填补和短时交通流预测方法研究[D].北京交通大学,2019.