基于数据挖掘的交通拥堵预测研究

张群慧[1]，彭　辉[1]，刘军华[2]
（1.湖南科技职业学院，湖南长沙 410004；2.湖南邮电职业技术学院，湖南长沙 410015）

【摘要】大众出行需求增高，传统的城市交通已不能满足人们出行的各种需求，在数据挖掘技术快速发展的背景下，智慧交通应运而生。作为智慧交通的一种典型应用，交通拥堵预测越来越受大众的青睐。探讨数据挖掘技术在智慧交通道路拥堵预测案例中的应用，分别从数据采集、数据清洗和预处理、模型构建与应用等方面进行说明。

【关键词】智慧交通；数据挖掘；聚类分析；拥堵预测
【doi:10.3969/j.issn.2095-7661.2022.03.009】
【中图分类号】TP311.13;U491.265　　【文献标识码】A　　【文章编号】2095-7661(2022)03-0036-04

Research on Traffic Congestion Prediction Based on Data Mining

ZHANG Qun-hui[1], PENG Hui[1], LIU Jun-hua[2]
(1.Hunan Vocational College of Science and Technology, Changsha, Hunan, China 410004;
2.Hunan Post and Telecommunication College, Changsha, Hunan, China 410015)

Abstract: With the increase of people's travel demand, traditional urban transportation can't meet people's travel needs. Under the background of the rapid development data mining technology, intelligent transportation came into being. As a typical application of intelligent transportation, traffic congestion prediction is increasingly favored by the public. This paper discusses the application of data mining technology in intelligent traffic road congestion prediction, and explains it from the aspects of data acquisition, data cleaning and preprocessing, and model construction and application.

Keywords: intelligent transportation; data mining; cluster analysis; congestion prediction

智慧交通是指在大数据技术快速发展的基础上，以互联网、物联网为途径对交通信息进行采集、分析和处理后构建成的交通运输服务体系[1]。它在智能交通的基础上，通过引入数据模型、数据挖掘相关技术，处理各种交通信息数据，为车辆调度、车辆管理、道路管理、交通控制及乘客出行提供智慧化服务。

1 智慧交通系统结构

智慧交通是一个庞大的系统，其基础设施包括交通道路上的车辆及其附属装备、监控设备等。智慧交通系统采集城市道路车辆运行轨迹、车辆运行状态数据，将城市运行车辆实时状态数据发送到地面大数据中心；地面大数据中心利用大数据、人工智能、数据挖掘技术对数据进行加工过滤、统计分析处理，为智慧应用提供数据支撑。

智慧交通系统架构由数据采集系统、数据平台和交通智慧应用组成。数据采集系统获取的数据是整个智慧交通系统的数据来源，该层数据包括传感器数据、车载导航数据、车载设备运行数据、公共交通场站电子站牌数据、交通监控数据等。智慧交通系统将数据采集层的数据整合发送到数据平台，数据平台对接收到的数据进行数据清洗与预处理（去除无效数据，将同一车辆的信息进行整合）、存储，然后基于清洗后的数据构建智能模型，利用模型处理的结果为智能交通应用提供智能支持。智慧交通的系统架构如图1所示。

[收稿日期] 2022-06-20
[作者简介] 张群慧（1973—），男，湖南长沙人，湖南科技职业学院副教授，高级工程师，研究方向：计算机技术、智能控制、模式识别及算法。
[基金项目] 2019年湖南省社会科学成果评审委员会课题“大数据时代智慧交通应用与发展策略研究”（课题编号：XSP19YBC189）。

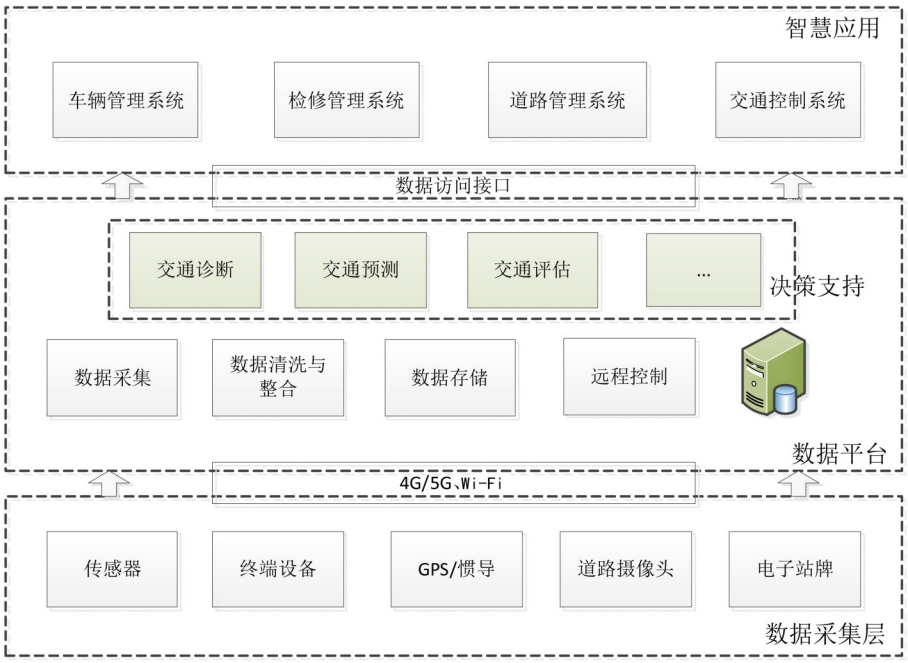

图1　智慧交通架构示意图

在智慧交通体系中，智慧应用直接服务于乘客用户，有很大的需求空间。智慧应用包含诸多场景，例如，使用图像识别技术对交通违法行为进行监测，识别出违法车辆具体违法行为；利用深度学习模型对轨道车辆状态进行诊断和评估，指导轨道交通检修人员对车辆的检查和维修；利用数据挖掘技术对道路拥堵情况进行统计分析，提供道路通行情况给交管部门工作人员，以便对指定交通路段进行交通管控，并可结合导航系统指导司乘人员调整行车计划。

本文研究数据挖掘技术在智能交通数据平台中的应用，通过对交通车辆数据采集、数据清洗和预处理、模型构建和应用，对当前的城市交通拥堵情况进行诊断和预测，更好地服务智慧交通应用。

2 基于数据挖掘的交通拥堵预测模型构建

发生交通拥堵的原因很多，一般来说，可将其分为常发性和偶发性拥堵[2]。由于偶发性拥堵在数据表现上不具明显特征，故本文所描述的拥堵是指常发性的交通拥堵。本文探讨对城市交通拥堵情况进行诊断和预测，首先需要先对采集的数据进行预处理得到清洗后的数据集，然后选择并构建数据挖掘模型，最后根据模型输出的结果对交通情况进行预测。

2.1 数据采集

对交通数据进行挖掘，首先要对所统计分析的车辆及附属设备、监控设备等进行数据采集。数据平台需要采集多种类型设备的数据，比如，车载GPS数据、传感器数据、Modbus协议数据、MVB接口数据、文件数据等。

车载数据采集使用系统日志采集法，由车载监控设备收集所在车辆信息后，发送到地面数据平台。同一辆车可能存在多种不同的数据：例如一些基础数据、告警数据和故障数据需要有可靠的网络来进行数据传输，如果数据有丢失需要重新传递，针对这部分数据使用TCP传输通道传递数据；而一般性的运营数据，在数据丢失的情况下，不需要考虑重传，针对这部分数据可以使用UDP传输通道传递数据。对于大文件数据（例如视频文件）传输，则考虑使用FTP传输通道进行文件上传。车辆监控设备与地面数据平台传输的接口设计应尽量简单、通用，具体架构如图2所示。

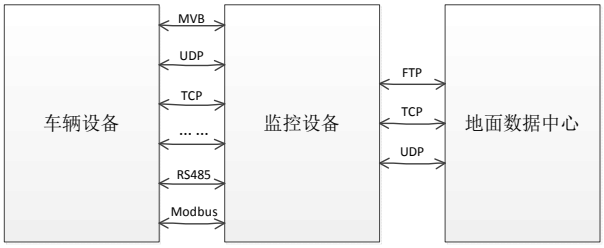

图2　数据采集架构示意图

监控设备可以将所在车辆的传感器数据、GPS导航数据、Modbus或MVB协议接口数据封装成一个数据包，通过UDP传输协议上报给数据平台。对于列车的故障数据，则应使用TCP协议，以确保关键数据的传输。视频数据和故障文件的传输则使用FTP传输通道上传。

2.2 数据清洗和预处理

数据清洗和预处理是大数据应用的一个重要

处理环节。本文研究道路拥堵预测模型，模型的构建依赖特定时间段车辆运营状态数据，包括数据记录时间、当前时刻车辆位置、车辆速度、载运数量以及当时的环境温度、湿度等。而另外一些与车辆运营状态无关的数据需要在构建模型之前，对其做删除处理。

考虑到数据采集设备上传的数据存在数据包重复、数据丢失等问题，在数据平台层面，需要对其进行数据的整合（例如，数据记录去重、删掉缺失值和异常值的记录）。不同车辆上传的数据格式存在差异，为方便应用，需要对数据格式进行统一格式化处理。

2.3 预测模型的分析与选择

数据挖掘技术是指在没有明确假设的前提下，通过计算机相关技术从大量数据中搜索隐藏的信息，发现知识。数据挖掘的主要任务包括关联分析、聚类分析、分类和预测等。其中关联分析的前提是多个对象间存在某种关联关系，通过关联着的一个对象状态能推导出所关联其他对象的状态。聚类分析是一种无监督的机器学习方法，其核心思想是“物以类聚，人以群分”，即把不同样本分割为有较多相似性子集合的过程。分类算法属于有监督的机器学习方法，分类算法的作用就是为当前样本寻找一个合适的类别表示，分类的规则来自于对历史经验数据集的分析。预测是指利用以往数据的变化规律，对未来变化趋势的一种预测。数据挖掘技术和机器学习技术，在分类算法、聚类算法上高度重叠，为此，数据挖掘任务可以充分使用机器学习体系完整的理论和方法来构建数据识别模型，从数据中挖掘出有价值的信息。本文使用聚类算法来实现对交通拥堵的预测。聚类算法的核心代表有K均值聚类算法和DBSCAN算法。

K均值聚类算法有一个核心参数——类中心点个数N，该参数值由算法的使用者给出。根据类中心点参数N，算法会随机生成N个中心点，并以这些中心点为基准，计算其他样本到各中心点的距离，并以距离的远近作为类别划分的依据，将每个样本归于不同的聚类簇。以上过程不断迭代，可将样本记录从一个组移动到另一个组来进行类别划分。

DBSCAN聚类算法使用“邻域”参数(ε,MinPts)来表达样本间的紧密性。对于任意样本点，分布于该样本点半径ε区域内的所有样本点的数量就是该点的密度。DBSCAN聚类算法基于样本点之间存在的密度直达、密度可达和密度相连关系[3-4]，对样本进行聚类划分。DBSCAN聚类过程首先是寻找样本中的所有核心点，待所有核心点被选出后，再以每个核心点为参考点将所有与之有密度关联的点划为一个关联区域。最后将所有关联的区域进行整合，得到最后的聚类结果。

K均值聚类和DBSCAN算法都能对样本集进行聚类划分，它们的区别在于：使用K均值聚类算法需要用户根据样本的大致分布情况设置对应的K值，而DBSCAN算法不需要用户做特定的设置；使用K均值聚类算法对球形分布的样本聚类效果良好，而DBSCAN算法则对样本的分布情况不敏感，它对任意形状分布的样本均能达到较好的聚类效果。从聚类的结果上看，DBSCAN算法能够找出样本中的离群点或异常点，而K均值聚类算法则不能找出异常点。

本文对城市道路交通进行拥堵预测，需要先对城市主要道路上的车辆按照其行驶速度进行等级划分，根据等级划分的结果，然后再对同一等级内的所有车辆按照区域位置进行聚类，从而得到每种拥堵情况发生的具体区域。为此，本文构建的模型需要进行两次聚类，由于第一次聚类涉及的通行等级是固定设置好的，所以使用K均值聚类算法比较合适，而第二次聚类的结果需要去除异常的样本点，故使用DBSCAN算法更为合适。

2.4 模型构建

对数据进行预处理后，以所统计车辆的标识号、车辆当前位置、当前车速、单位时间内的通行距离作为数据集进行聚类划分，对道路拥堵情况进行判别和预测。具体做法如下：

1）以所统计车辆的当前车速、单位时间内通行距离来构造矢量空间，根据交通拥堵的级别数量设置K值，调用K均值聚类算法对车辆进行一次聚类划分，得到包含K个类别的聚类结果集合C={C1,C2,…,Ck}，集合C中的每一个元素分别表示某种拥堵级别车辆的集合，例如C1表示拥堵级别最为严重的车辆的集合。通过K均值聚类，每个样本都被划入到特定类C1,C2,…,Ck中。

第一次聚类使用K均值聚类算法，以交通拥堵的级别数量为K值，以表示拥堵等级的速度阈值作为中心点center，来对车辆速度数据data进行一次聚类划分，核心算法如下：

```
01:for i in range(n):
02:for j in range(k):
03:dist[i, j] = np.sqrt(sum((data[i, :] - center[j, :]) **2))
```

04:dist[i, k] = np.argmin(dist[i, :k])

第一次聚类所得到的结果即表示每辆车的行驶缓慢程度，它仅表明单辆车的行驶状况，不能反映真实的拥堵情况。例如，司机有意放缓了行驶速度，而所行驶的道路并不拥堵。为此，还需要进行第二次聚类。

2）对上述集合C中的每一个簇内样本进行二次聚类。为方便理解，以第i个簇Ci为例进行描述：使用Ci簇内样本所对应的车辆位置信息（经、纬度数据）作为特征进行DBSCAN密度聚类，将Ci中的样本划分为n个不同的子类别，得到基于Ci的第二次聚类结果集合Si={Ci1,Ci2,…,Cin}，Si中的每一个元素Cij分别表示某区域范围内达到了特定拥堵级别的车辆集合。对聚类结果C中所有簇进行二次聚类后得到最终的结果S={S1,S2,…,Sk}。

第二次聚类使用DBSCAN密度算法，以车辆间的距离gap为密度，以区域车辆的数量count为最小样本数量，对每个簇Ci内的车辆进行第二次聚类划分，利用Sklearn开发库实现的代码如下：

```
01: esp_dbscan=DBSCAN(eps=gap,minsamples=count)
02: esp_dbscan.fit(x)
03: esp_dbscan.fit_predict(x)
```

第二次聚类所得到的聚类结果S表现为同样行驶缓慢程度的车辆集中发生的区域。

3 拥堵预测模型的应用

利用聚类模型对当前交通数据进行聚类分析后，得到第一次聚类簇C和第二次聚类簇S，通过对聚类结果C和S的分析，即可实现对道路拥堵情况的预测。

由于单台车行驶缓慢并不意味着当前所在区域道路发生了拥堵，此时有可能是司机单方面的驾驶行为导致，为此，需结合同区域同样行驶缓慢程度车辆的数量进行判断，如果同样行驶缓慢车辆的数量大于指定的阈值，则判定该区域发生了拥堵，否则认为该区域通行正常。

先遍历S集合中的元素{S1,S2,…,Sk}，再对每个Si所包含的聚类簇{Ci1,Ci2,…,Cin}进行遍历处理：统计簇Cij中的样本数量，根据样本数据的多少决定是否是有效的聚类簇。示例代码如下：

```
01：C={C1,C2,…,Ck}，其中每个Ci由{Ci1,Ci2,…,Cin}聚类簇组成
02：for i in range(len(C)):
03：for j in range(len(C[i])):
04：count = len(C[i][j])
05：if count >= 阈值:
06：当前簇内样本车辆所处区域发生了等级为i的拥堵
```

通过两层循环遍历S集合中的元素Cij，可以对Cij的样本数量进行判断，如果特定区域样本数量大于指定阈值，则表明Cij样本所在的区域发生了等级为i的拥堵。

通过使用上述数据挖掘技术，对测试数据进行聚类划分后，即可根据聚类划分情况预测当前城市道路的拥堵情况。但城市道路交通瞬息万变，为此，若要对交通拥堵情况进行实时预测，系统需每隔一段时间就进行一次聚类分析，再根据聚类分析的结果，对交通情况进行预测。当然，要实现对特定区域交通拥堵的精确预测，需要调用专业的地图服务接口程序，将拥堵区域映射到地图上，实现对城市道路的交通预测。

4 结语

数据挖掘技术在智慧交通中的应用包括车辆故障诊断、故障预测、交通拥堵判断与预测等诸多方面。提出了一种使用K均值聚类和DBSCAN聚类算法对交通数据进行拥堵预测的方案设计，但由于交通系统的复杂性，若要将方案落实到实际的交通系统中，需要加大数据采集层的软硬件系统的投入，同时需要考虑大规模数据采集存在的高时延问题，以及数据平台运行的高负荷及高开销等问题[5]。

【参考文献】

[1]王洪斌.大数据背景下人工智能在智慧交通中的应用研究[J].电脑知识与技术,2021(12):198-199.

[2]林立春,洪东,刘华.基于大数据分析和多模型融合的交通拥堵高效预测技术[J].西部交通科技,2021(7):151-155.

[3]吴国强,姚建锋,管敏渊,朱雪松,吴凯,李浩言,宋珊珊.基于DBSCAN的电力变压器故障诊断[J].武汉大学学报(工学版),2021(12):1172-1179.

[4]翟玉健,余承智,高星.一种基于DBSCAN聚类的雷达点迹处理方法[J].舰船电子对抗,2021(5):58-61.

[5]吴建波,朱文霞,剧亮,许致芳.边缘计算在智慧交通系统中的应用[J].计算机与现代化,2021(12):103-109.