第 44 卷　第 4 期
2020 年 8 月

武汉理工大学学报(交通科学与工程版)
Journal of Wuhan University of Technology
(Transportation Science & Engineering)

Vol. 44　No. 4
Aug. 2020

基于 Spark 的路网交通运行分析系统*

杨　孟[1)]　许宏科[1)]　钱　超[1)]　朱　熹[2)]

(长安大学电子与控制工程学院[1)]　西安　710064)　(深圳市城市交通规划设计研究中心有限公司[2)]　深圳　518021)

摘要:应用 Spark 大数据平台,设计了一种路网交通运行分析系统. 以实时交通流数据为基础,结合 k-means 和随机森林算法构建了路网交通运行态势判别模型,选择分布式文件系统存储和弹性分布式数据集并行处理交通数据,实现了路网交通运行状态的实时判别. 利用西奥克兰地区路网车检器数据展开实例分析,结果表明:该系统的运行速度快;分布式 k-means 算法相较于传统的算法聚类结果的平均相对误差约为 7.3%,具有较高的准确性;分布式随机森林算法的精确度、召回率和 F 度量分别为 98.98%,98.96%,98.97%,分类结果均优于逻辑回归、多层感知器算法.

关键词:公路运输;交通运行分析;数据挖掘;大数据平台;Spark;机器学习

中图法分类号:U495　　**doi**:10.3963/j.issn.2095-3844.2020.04.023

0　引　　言

随着智能交通系统(intelligent transportation system, ITS)研究的深入展开,道路交通数据规模和复杂度呈爆发式增长,呈现出大数据的"6V"特征[1]. 采用传统的串行处理方式,其计算速度已无法满足大数据环境下实时业务需求. 因此,采用并行化与分布式的数据处理技术来提高交通信息处理水平成为当前交通大数据平台研究的热点. 建立综合运输服务大数据平台,促进交通运输大数据产业化应用成为迫切的行业需求[2].

目前,国内外在交通大数据应用领域积极展开相关研究. 由于传统的数据存储方法无法解决海量交通数据的高效存储和快速增长问题,Zhu 等[3-4]采用 Hadoop 的分布式文件系统进行交通数据并行存储,并应用 MapReduce 分布式计算框架实现对交通数据的统计分析,提高了海量交通数据的存储和计算效率;Rathore 等[5]根据城市交通监控数据,利用 Hadoop 的 MapReduce 机制,在并行环境下将视频进行分块处理,提高城市道路违规检测的效率. 为提高交通数据处理能力,孙卫真等[6]改进了分布式调度算法模型,优化了 Hadoop 的处理能力,从而弥补了传统调度算法实时性的不足;Park 等[7]采用数据挖掘算法对交通流数据进行聚类与分类分析,提出了一种改进的交通事故预测模型;Fan 等[8]以 ETC 收费数据为基础,采用随机森林算法与 Hadoop 构建大数据机器学习分析平台,实现公路旅行时间的预测;Chen 等[9]利用历史的交通速度,在 Hadoop 平台上集成了 KNN 算法和高斯过程对道路速度进行预测. 以上研究主要利用历史数据对交通数据进行处理,为了实现交通数据的实时处理,Tsai 等[10]利用 Spark 平台实时处理数据的能力,构建了一种可以实时提供路网交通量的系统;黄廷辉等[11]利用道路检测数据,提出了一种分布式城市交通流预测模型,实现了实时、准确的交通流预测;段宗涛等[12]在 Hadoop 平台上设计并实现了一种多路径的实时交通流分配方法,解决了传统交通分配方法的难以保证交通流均衡性问题;陈钊正等[13]结合实际的交通流数据,利用聚类算法对交通流量和速度进行聚类分析,给定了交通状态划分方法,结果反映了实时、准确交通运行状态.

收稿日期:2020-07-05

杨孟(1994—):男,硕士生,主要研究方向为数据挖掘、交通数据分析

* 中央高校基本科研业务费专项资金项目资助(310832161006,310821173102)

综上所述,应用分布式系统进行交通大数据研究集中在对传统交通模型的改进以及对交通信息预测,缺少利用实时的交通数据对路网交通运行状态进行更加合理、准确的分析研究,从而进行多指标综合评价. 因此,本文设计了一种基于 Spark 的路网交通运行状态分析系统,以实时的交通流指标为基础,实现对路网运行状态的判别. 结合真实路网交通数据,对系统分析结果进行综合评价,验证了系统的准确性与高效性.

1 路网大数据机器学习平台

Spark 是 Apache 项目的一个开源集群运算框架[14],具有分布式存储和并行计算的能力,同时还提供了机器学习算法编程的接口,以及利于迭代运算的并行化执行机制,保证平台在可接受的时间内完成大规模数据的学习和训练. 本文采用 Spark 技术搭建的路网大数据机器学习平台总体框架,见图 1. 在 Linux 操作系统上搭建 Hadoop 与 Spark 平台,利用 Hadoop 平台的分布式文件系统(hadoop distributed file system, HDFS)作为路网大数据机器学习平台的数据存储层,负责底层交通数据存储管理. 数据处理层利用 Spark SQL 对交通数据进行读取与查询,并将读取的结果作为 SparkR 的输入,利用 Spark 调用的 k-means 算法和随机森林算法实现路网交通运行状态的判别;并在数据应用层对数据判别结果进行研究分析. 在数据的存储、处理与应用过程中,由于 Spark 平台的独立调度器(standalone)模式较为简单方便,无需依赖其他任何的资源管理系统,利用 Standalone 模式实现底层资源调度;同时,利用弹性分布式数据集(resilient distributed datasets, RDD)进行交通数据处理任务的并行执行;相较于 MapReduce 方法,RDD 利用高速内存代替了低速磁盘 I/O 操作,提高了整体的运算效率.

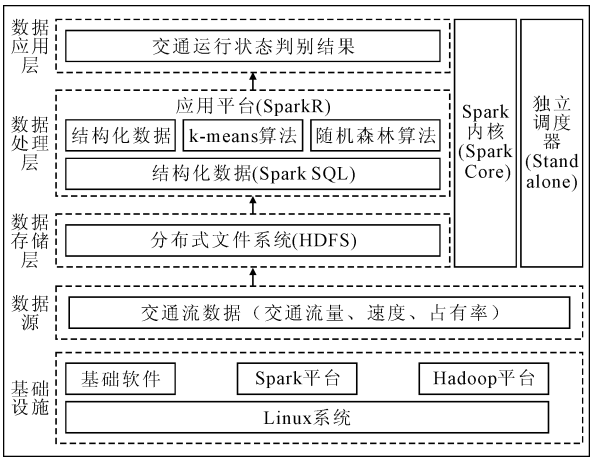

图 1 路网大数据机器学习平台

2 路网交通运行状态研判

2.1 路网交通运行状态聚类

路网畅通程度是描述道路运行状态的重要指标. 2012 年,交通运输部在《公路网运行监测与服务暂行技术要求》中以路段平均车速为标准,将道路交通运行状态划分为“畅通”“基本畅通”“轻度拥堵”“中度拥堵”“严重拥堵”五级. 但不同的道路速度可能会有多种不同的道路状况,仅用平均车速来度量交通运行状态缺乏科学性和可靠性. 因此,本文以道路交通流量、速度和占有率作为评价交通运行状态的指标,采用聚类算法将道路拥堵程度划分为五种状态. 传统的 k-means 算法由于其原理简单而被广泛使用,当数据量较大时,算法的时间开销非常大. 本文采用分布式 k-means 算法进行聚类分析,将大量交通流数据划分为多块子数据,采用多个处理器并行计算,从而减少算法的运算时间. 分布式 k-means 算法的基本思想基本过程如下.

1) 从高速公路交通流数据集 $D=\{x_1,x_2,\cdots,x_n\}$ 中,随机选择 k 个中心点 m_j,并将其存入文件 clusterList 中.

2) 在路网大数据机器学习平台的分布式文件系统中,每个节点都包含部分数据集 D_i,将文件 clusterList 分发给分布式文件系统的每个节点中.

3) 在每个子数据集 D_i 中,计算非中心数据 x_i 到 k 个中心点数据 m_j 的距离 $d(x_i,m_j)$,如果 $d(x_i,m_j)=\min\{d(x_i,m_j),i=1,2,\cdots,n';j=1,2,\cdots,k\}$,则将 x_i 划分到中心数据 m_j 的类中.

4) 直到每个节点上非中心数据划分到 k 个聚类以后,每个节点上都形成 k 个簇,综合所有节点上具有相同中心点的簇,使其成为 k 个新簇;分别计算综合后 k 个新簇的均值作为新中心数据 m_j^*,并将新的中心点保存在 clusterList 文件中.

5) 计算 k-means 算法的误差平方和准则函数 J,若聚类准则函数收敛或聚类迭代达到最大,则得到最终聚类结果;否则重复步骤 2)、3)、4)继续迭代,直到满足聚类停止条件.

6) 迭代结束,得到交通流运行状态聚类结果.

2.2　路网交通运行状态判别

利用 k-means 算法实现路网交通运行状态聚类后，每条交通流数据被赋予一个特定的分类标签，其聚类结果为 $T=\{(x_i,m_j);i=1,2,\cdots,n;j=1,2,\cdots,5\}$. 其中：$x_i$ 为交通流运行数据，包括交通流量、速度和占有率，n 为数据集记录数，m_j 表示交通流运行数据聚类后的标记即五种交通运行状态. 随机森林算法(random forest，RF)是以聚类产生的类别标签为规则，判别数据与分类规则之间的关系. 将带标签的交通流数据作为随机森林算法的输入数据，实现路网运行状态判别，其具体判别过程如下.

1) 以高速公路交通流运行数据集 $D=\{x_1,x_2,\cdots,x_n\}$ 与各样本对应的客户类别为基础，采用 Bootstrap 重采样技术从数据集 D 中有放回地随机抽取 numTrees 个子数据集，并将 numTrees 个子数据集 D_i 基本均匀的分配到路网大数据机器学习平台的所有节点中.

2) 分别从平台所有节点的数据集 D_i 中随机选取 $M(M\leqslant 3)$ 个特征属性，将 M 个特征属性作为数据集 D_i 的特征属性.

3) 并行训练所有节点的数据集 D_i，以计算信息增益的方式确定最优的属性划分点，构建 numTrees 棵交通流运行状态判别决策树.

4) 利用 numTrees 棵决策树形成交通流运行状态判别随机森林，并综合 numTrees 棵决策树的判别结果，按 numTrees 棵树分类器投票决定最终分类结果.

2.3　路网交通运行状态判别结果评价

综合评价路网判别结果，本文引入交通运行状态混淆矩阵见表 1，其中，每一列代表了交通运行状态的类别，每一行代表了交通数据真正的归属类别. 混淆矩阵可以直观反应实际交通运行状态与判别结果的分布情况，根据混淆矩阵提取出精确度、召回率和 F 度量等指标来评判判别结果的准确性.

表 1　交通运行状态混淆矩阵

交通状态	交通运行状态判别结果					合计
	畅通	基本畅通	轻度拥堵	中度拥堵	严重拥堵	
畅通	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	P_1
基本畅通	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	P_2
轻度拥堵	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	P_3
中度拥堵	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	P_4
严重拥堵	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	P_5
合计	P'_1	P'_2	P'_3	P'_4	P'_5	M

1) 精确度 $Prec$　描述交通运行状态判别结果正确的百分比，其中 P'_j 为判别交通运行状态为 j 的测试数据记录数.

$$Prec_j=\frac{n_{jj}}{P'_j}\times 100\% \tag{1}$$

2) 召回率 Rec　描述交通运行状态判别模型中正确结果占实际交通运行状态的百分比，其中 P_j 为实际交通运行状态为 j 的测试数据记录数.

$$Rec_j=\frac{n_{jj}}{P_j}\times 100\% \tag{2}$$

3) F 度量　精确度 $Prec$ 与召回率 Rec 的调和均值，体现了判别模型的稳定性.

$$F_j=\frac{2\times Rec_j\times Prec_j}{Rec_j+Prec_j}\times 100\% \tag{3}$$

3　实例分析

3.1　实验平台搭建

PeMS(performance measurement system)是美国加州运输局运行监测系统，包含近 40 000 个检测器的实时路网交通数据. 本文选取西奥克兰(West Oakland)地区高速公路作为实验路网，包括 I880 号、I580 号、I980 号、I80 号和 SR24 号高速公路，共布设 57 个车辆检测器，实验路网见图 2. 以 2016 年 5 月 29 日—9 月 3 日的交通流运行数据作为基础数据，具体数据量为 1 608 768 条，采样间隔为 5 min. 实验路网交通流运行参数见表 2.

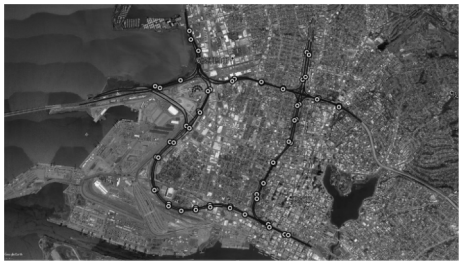

图 2　实验路网

表 2　交通流运行原始数据表

检测器编号	时间	交通流量/(veh·5 min^{-1})	时间占有率/%	速度/(mile·h^{-1})
408000	2016-05-29 T00:00	113	1.7	68.1
408000	2016-05-29 T00:05	105	1.6	67.5
408000	2016-05-29 T00:10	82	1.2	68.0
408000	2016-05-29 T00:15	73	1.0	68.1
⋮	⋮	⋮	⋮	⋮

本文利用 5 台 PC 机搭建包含一个控制节点和四个计算节点的路网大数据机器学习平台，处理器 Intel(R) Core(TM)2 i5-6500@3.20 GHz，4 G内存. 在路网大数据机器学习平台中的所有节点上均安装有 Linux(ubuntu 12.04)操作系

统,并配置 Spark 所需的软件,包括:Java,Hadoop,Scala,Spark 和 R.

3.2 聚类可靠性分析

聚类结果的可靠性决定了路网运行分析系统准确性.因此,本文通过对比并行化聚类算法和传统的聚类算法结果、并行化聚类算法结果和实际交通特性,对聚类结果进行评价.

3.2.1 并行化聚类和传统的聚类结果分析

相较于传统的聚类算法,路网大数据机器学习平台对预处理后的交通流数据进行并行计算,大幅度提高了聚类效率,聚类结果统计见表 3.由表 3 可知,两种聚类方式的聚类结果占比基本一致,其平均相对误差约为 7.3%,说明并行化聚类和传统聚类算法结果具有一致性.

表 3 并行化聚类与传统聚类结果

聚类方式	类别	畅通	基本畅通	轻度拥堵	中度拥堵	严重拥堵
传统聚类	样本数目	642 922	502 614	214 788	169 460	72 076
	所占比例/%	40.14	31.38	13.41	10.58	4.50
并行化聚类	样本数目	649 489	534 808	213 507	132 781	71 115
	所占比例/%	40.55	33.39	13.33	8.29	4.44

3.2.2 并行化聚类结果时间特性分析

图 3 为 401416 号检测器 6 月 7 日并行聚类结果时间分布特性图,采用"1""2""3""4""5"表示交通运行状态的"畅通""基本畅通""轻度拥堵""中度拥堵""严重拥堵".由于 I980 号高速公路具有早晚高峰特点,而 401416 号检测器处于 I980 号高速公路下行线上.由图可知:401416 号检测器并行聚类结果时间分布特性具有早高峰特点,在早晨 08:00 前后道路交通量和占有率达到最高,同时道路上车辆的速度下降到最低值,与交通流运行特性是一致的,说明交通流运行数据并行聚类结果是可靠的.

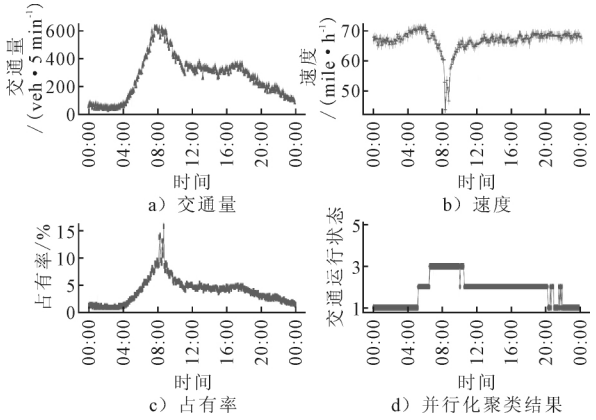

图 3 401416 号检测器样本与并行化聚类结果时间分布

3.3 判别准确性分析

3.3.1 传统判别与并行化判别结果评价

在单机和路网大数据机器学习平台上分别构建路网交通运行状态判别模型,其中 85%的路网数据作为训练集,15%的路网数据作为测试集,构建传统判别与并行化判别结果的交通运行状态混淆矩阵,并从混淆矩阵中提取出两种判别结果平均精确度、召回率和 F 度量见图 4.由图 4 可知,在路网大数据机器学习平台上并行化判别路网运行状态会影响其判别结果,并行化判别的精确度、召回率和 F 度量略低于传统判别,但均达到 98.5%以上,说明并行化判别结果的准确性依然可靠.

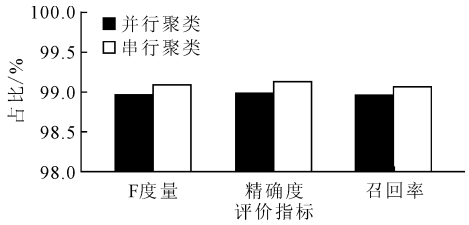

图 4 传统判别与并行化判别结果评价指标平均值对比

3.3.2 并行化判别模型评价

在路网大数据机器学习平台中,为评价判别模型的准确性,本文选用逻辑回归模型(Logit)、多层感知器(multi-layer perception, MLP)和随机森林算法(RF)进行对比.采用 85%数据作为训练集和 15%数据作为测试集进行实验,图 5 为在不同交通运行状态下 Logit,MLP,RF 分类算法的精确度、召回率和 F 度量对比,图 6 为 Logit,MLP,RF 分类算法的平均精确度、召回率和 F 度量对比.由图 6 可知,随机森林算法的精确度、召回率和 F 度量高于 Logit 和 MLP 算法,并均达到 98%以上,说明在路网大数据机器学习平台中,随机森林算法的准确性相较于其他分类算法准确性较高.

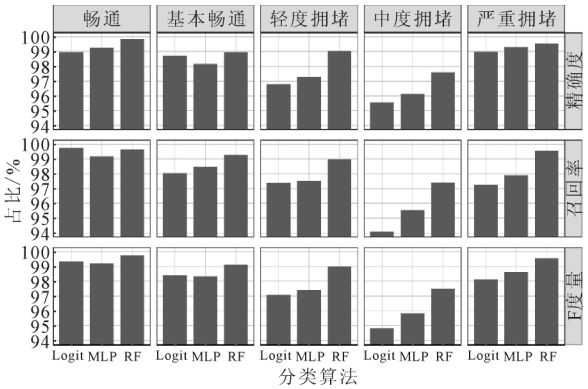

图 5 不同分类算法下五种判别结果的指标对比

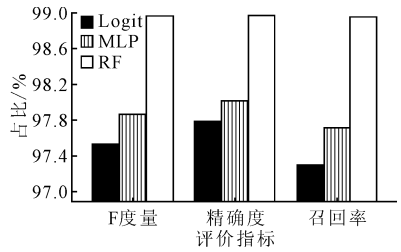

图 6　不同分类算法下判别结果的指标对比

3.4　系统性能分析

3.4.1　运行时间

以路网交通流数据为基础，不断增加数据规模，分析在不同计算节点数下路网交通运行分析系统的运行时间，结果见图 7. 当数据规模较小时，增加计算节点的数量对系统的运行时间影响不大；随着数据规模的增大，系统中计算节点的数量越多，其运行时间的越短.

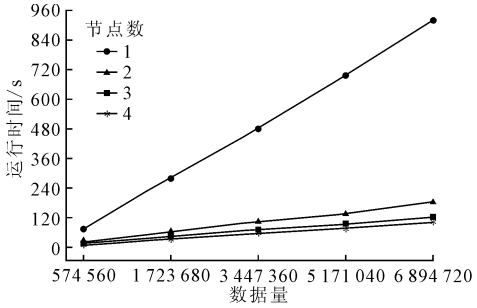

图 7　不同节点运行时间对比

3.4.2　加速比

在在不同数据规模下，改变系统中计算节点的数量，分析并行判别系统的加速比，结果见图 8. 增加计算节点的数量，加速比均会上升；当数据规模较少时，加速比随着计算节点数量的增加先增大后趋于平稳；当数据规模较大时，增加系统中计算节点，系统的加速比也不断上升.

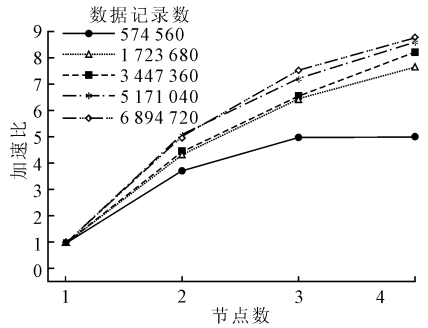

图 8　路网交通运行分析系统加速比

3.4.3　可扩展性

以不同规模数据为基础，通过改变系统节点数量分析路网交通运行分析系统的运行时间，结果见图 9. 增加系统中计算节点的数量，数据的运行时间均有所下降；数据规模越大，系统的运行时间下降的幅度越大，说明路网交通运行分析系统适用于不同规模数据处理，具有良好的可扩展性.

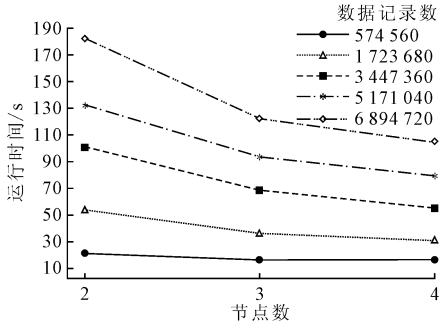

图 9　路网交通运行分析系统可扩展性

4　结　　论

1）以交通流量、速度和占有率为基础进行交通运行状态评价，从而更加全面、准确的反映路网中的路网运行状态.

2）经实验证明，相较于传统的运算系统，本文提出的并行运算系统结果依然可靠、准确，系统的加速比提升了近 50%，并具有良好的可扩展性，能更有效的对大规模数据进行处理.

3）本文采用定点检测器采集的交通流量、速度和占有率数据实现路网交通状态的判别，检测器的布设密度对实际结果具有一定的影响. 因此，在未来的研究中，采用定点检测器数据与动态采集设备的数据相融合，能进一步提高交通状态判别的准确性和可靠度.

参 考 文 献

[1] 马莹莹，邹祥莉，徐建闽. 基于宏观基本图的路网交通拥堵甄别方法研究[J]. 武汉理工大学学报(交通科学与工程版)，2019，43(4)：575-579.

[2] 徐建闽，邹磊. 基于交通拥堵指数的禁左交通组织研究[J]. 武汉理工大学学报(交通科学与工程版)，2020，44(2)：201-205.

[3] XIA Y，CHEN J，LU X，et al. Big traffic data processing framework for intelligent monitoring and recording systems[J]. Neurocomputing，2016(1)：55-59.

[4] ZHU L，YUN L. Distributed storage and analysis of massive urban road traffic flow data based on hadoop：web information system & application conference(WISA)[C]. IEEE，London，2015.

[5] RATHORE M M，SON H，AHMAD A，et al. Re-

al-time video processing for traffic control in smart city using Hadoop ecosystem with GPUs[J]. Soft Computing, 2018,22(5):1533-1544.

[6] 孙卫真,王秀锦,徐远超. 交通信息分布式处理中的Hadoop调度算法优化[J]. 计算机工程与设计,2014(4):1269-1273.

[7] PARK S, KIM S, HA Y. Highway traffic accident prediction using VDS big data analysis[J]. The Journal of Supercomputing, 2016,72(7):2815-2831.

[8] FAN S S, SU C, NIEN H, et al. Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection[J]. Soft Computing, 2018,22(17):5707-5718.

[9] CHEN X, PAO H, LEE Y. Efficient traffic speed forecasting based on massive heterogenous historical data[C]. IEEE International Conference on Big Data, Beijing,2014.

[10] TSAI J, CHANG T, FANG Y, et al. A real-time traffic flow prediction system for national freeways based on the spark streaming technique[C]. IEEE International Conference on Consumer Electronics, Taiwan,2018.

[11] 黄廷辉,王玉良,汪振,等. 基于Spark的分布式交通流数据预测系统[J]. 计算机应用研究,2018(2):405-409.

[12] 段宗涛,李莹,郑西彬,等. 基于Hadoop平台的实时多路径交通流分配算法[J]. 中国公路学报,2014(9):98-104.

[13] 陈钊正,吴聪. 多变量聚类分析的高速公路交通流状态实时评估[J]. 交通运输系统工程与信息,2018(3):225-233.

[14] 马莹莹,邹祥莉,徐建闽. 基于宏观基本图的路网交通拥堵甄别方法研究[J]. 武汉理工大学学报(交通科学与工程版),2019,43(4):575-579.

Spark-based Operational Analysis System for the Expressway Network

YANG Meng[1] **XU Hongke**[1] **QIAN Chao**[1] **ZHU Xi**[2]

(*School of Electronic and Control Engineering, Chang'an University, Xi'an* 710064, *China*)[1]

(*Shen Zhen Urban Transport Planning Center, Shenzhen* 518021, *China*)[2]

Abstract: Using Spark big data platform, the expressway network operational analysis system was designed. Based on the real-time traffic flow data, combined with k-means and stochastic forest algorithm, a road network traffic situation discrimination model wais constructed. The distributed file system and the resilient distributed dataset were selected to process the traffic data in parallel, realizing the real-time discriminant of the operation status for the expressway network. An example analysis was carried out by using the vehicle detector data of west Auckland road network. The results show that the system runs fast. Compared with the traditional clustering algorithm, the average relative error of the distributed k-means algorithm is about 7.3%, which has higher accuracy. The accuracy, recall rate and F-measure of distributed random forest algorithm are 98.98%, 98.96% and 98.97%, respectively, and the classification results are better than those of logistic regression and multilayer perceptron algorithm.

Key words: highway transportation; traffic operation analysis; data mining; big data platform; Spark; machine learning