ICS 35.240.60
CCS R 07

DB34

安　徽　省　地　方　标　准

DB34/T 4641—2023

交通管理大数据中心数据模型建设规范

Specification for business modeling of traffic management big data center

2023－12－18 发布　　2024－01－18 实施

安徽省市场监督管理局　　发 布

目　　次

前言 II
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　建设流程 1
5　数据处理 2
5.1　数据提取 2
5.2　数据治理 2
6　特征选择 3
6.1　构造衍生特征 3
6.2　特征转换 4
6.3　特征筛选 4
7　数据建模 5
7.1　业务规则类模型 5
7.2　预测预警类模型 6
7.3　异常检测类模型 7
8　模型评估 8
8.1　评估指标 8
8.2　评估方法 8
8.3　评估处理 9
9　模型发布 9
附录A（资料性）　驾驶人数据项 10
附录B（资料性）　机动车数据项 11
附录C（资料性）　违法数据项 12
附录D（资料性）　事故数据项 13
附录E（资料性）　事故人员数据项 15
附录F（资料性）　过车数据项 17
附录G（资料性）　道路数据项 18
附录H（资料性）　卡口设备数据项 19
附录I（资料性）　气象数据项 20
附录J（资料性）　其他数据项 21

前　言

本文件按照 GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由安徽省公安厅提出并归口。

本文件起草单位：安徽百诚慧通科技股份有限公司、安徽省公安厅交通警察总队科技处、合肥工业大学、安徽省智能交通协会、合肥工业大学设计院(集团)有限公司、安徽中汇规划勘测设计研究院股份有限公司。

本文件主要起草人：朱文佳、胡博、汪春、张豪、张宏燕、张卫华、任冉冉、颜鹏、余烨、胡长江、谢晓琳、徐龙、秦忱忱、吴磊、耿伟、胡恒、杜礼、乔文、陈珊珊、丁俊美。

交通管理大数据中心数据模型建设规范

1 范围

本文件确立了交通管理大数据中心数据模型建设流程，并规定了交通管理大数据中心数据模型建设的数据处理、特征选择、数据建模、模型评估、模型发布。

本文件适用于交通管理大数据中心数据模型建设。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

交通管理大数据中心 traffic management big data center

公安交通管理部门设立的管理交通安全大数据的机构。

4 建设流程

交通管理大数据中心数据模型建设流程见图1。

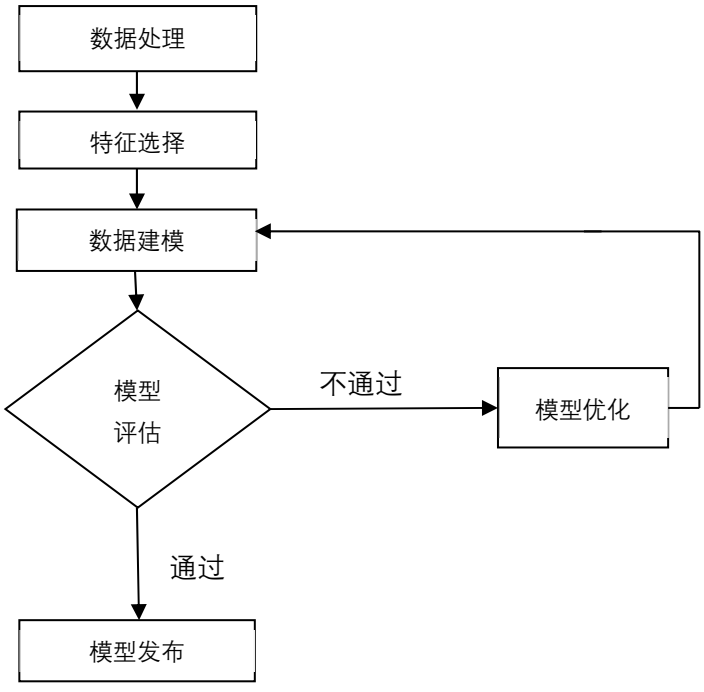

图1 交通管理大数据中心数据模型建设流程示意图

5 数据处理

5.1 数据提取

5.1.1 数据来源

交通管理大数据业务中心的数据来源包括但不限于：

——交通管理大数据综合应用平台，

——集成指挥平台，

——第三方外挂平台。

5.1.2 提取范围

5.1.2.1 交通管理大数据中心数据分为驾驶人、机动车、违法、事故、过车、道路、卡口设备、气象、其他数据。

5.1.2.2 驾驶人数据项见附录 A。

5.1.2.3 机动车数据项见附录 B。

5.1.2.4 违法数据项见附录 C。

5.1.2.5 事故数据项见附录 D、附录 E。

5.1.2.6 过车数据项见附录 F。

5.1.2.7 道路数据项见附件 G。

5.1.2.8 卡口设备数据项见附录 H。

5.1.2.9 气象数据项见附录 I。

5.1.2.10 其他数据项见附录 J。

5.1.3 提取方式

5.1.3.1 使用公安交通管理业务分布式汇聚管理平台采集 IoT 类型设备数据，支持相机 SDK 接入、GAT 1400 公安视图库标准协议接入、ftp 协议接入以及消息队列中间件数据接入等多种数据采集方式。

5.1.3.2 使用公安交通管理数智能力开放平台数据接入模块接入数据，支持离线数据集成接入、实时消息集成接入、服务接口集成接入等多种数据采集方式。

5.2 数据治理

5.2.1 非空数据核验

应对下列数据的非空数据进行核验：

——附录 A 中的身份证明号码、性别、初次领证日期，

——附录 B 中的号牌号码、号牌种类、车辆类型、使用性质、初次登记日期、身份证明号码、机动车状态、核定载客、检验报废期止、强制报废期止，

——附录 C 中的号牌号码、号牌种类、违法时间、违法行为、违法记分数，

——附录 D 中的事故编号、事故发生时间、死亡人数、受伤人数、号牌号码、号牌种类、是否逃逸，

——附录 F 中的号牌号码、号牌种类、经过时间，

——附录 G 中的道路代码、道路类型、道路名称、行政区划、管理部门，

——附录 H 中的设备编号、设备类型、使用状态、车道号、方向类型、点位编号，

——附录 I 中的设备编号、检测时段、检测时间，

——附录 J 中的身份证明号码。

5.2.2 重复数据去重

应对下列数据进行重复数据去重：

——附录 A 中的身份证明号码，
——附录 B 中的号牌号码、号牌种类，
——附录 C 中的违法编号，
——附录 D 中的事故编号，
——附录 G 中的道路代码，
——附录 H 中的设备编号，
——附录 I 中设备编号，
——附录 J 中的身份证明号码。

5.2.3 错误数据删除

应删除下列错误数据：

——附录 A 中的准驾车型代码不存在、身份证明号码长度不符合 18，
——附录 B 中的初次登记日期内容早于 2001 年，
——附录 C 中的单次违法记分数值为 1，3，6，12 以外的，
——附录 D 中的事故发生时间大于当前时间，
——附录 F 中的经过时间大于当前时间、号牌号码识别长度小于 7 或大于 8，
——附录 G 中的道路类型不存在，
——附录 H 中的设备编号不符合设计标准，
——附录 I 中的设备编号记录值为空。

5.2.4 规范数据类型

应对下列数据类型进行规范：

——数据类型不一致，如日期类型的数据实际却是字符或数字类型，应转成日期类型，
——附录 A 中下一清分日期、下一审验日期、初次领证日期、有效期始、有效期止、发证日期、出生日期，
——附录 B 中初次登记日期、最近定检日期、检验报废期止、强制报废期止、发行驶证日期、发登记证书日期、发合格证日期、保险终止日期，
——附录 C 中违法时间、处理时间、缴款日期、录入时间，
——附录 D 中开始侦查时间、结束侦查时间、事故发生时间、录入时间、更新时间，
——附录 F 中经过时间、录入时间，
——附录 I 中检测时间。

6 特征选择

6.1 构造衍生特征

6.1.1 附录 A 中驾驶人驾龄通过当前日期减去初次领证日期、驾驶人年龄通过当前日期减去出生日期；是否车型降级通过准驾车型和原准驾车型比较得到。

6.1.2 附录 B 中车辆是否强制报废通过当前日期减去强制报废期止、是否逾期未年检通过当前日期减

去检验有效期止、是否逾期未保险通过当前日期减去保险终止日期。

6.1.3　附录C中车辆违法次数通过对号牌号码和号牌种类分组计数得到、车辆总违法记分数通过对号牌号码和号牌种类分组求和得到、车辆有严重违法次数通过对违法类型做分组然后对号牌号码和和号牌种类计数得到。

6.1.4　附录D中车辆发生财产损失事故次数通过对号牌号码和号牌种类分组计数得到、车辆发生伤人事故次数通过对号牌号码和号牌种类分组计数得到、车辆发生亡人事故通过对号牌号码和号牌种类分组计数得到。

6.1.5　附录F中车辆过车天数通过号牌号码和号牌种类分组对日期进行计数得到、车辆经常经过点位通过对号牌号码和号牌种类与点位分组计数，然后再对号牌号码和号牌种类进行排序取计数最多的点位。

6.1.6　附录H中卡口在道路的位置通过公里数和米数相加得到。

6.2　特征转换

6.2.1　二值化法

将两个类别型的特征，转换成1、0。

如驾驶人数据中性别特征，转换成男性:1，女性:0；国籍转化为中国人:1，外国人:0；车辆数据中是否强制报废，是：1，否：0；是否逾期未年检，是：1，否：0；是否逾期未保险，是：1，否：0；事故数据中事故类型转成伤亡事故：1，财产损失事故：0。

6.2.2　哑变量法

将不能够定量处理的特征量化，对多类别型特征处理。

如车辆数据中车辆类型转换，大车：0001，小车：0010，摩托车：0100，其他车：1000；车辆数据中车辆使用性质，客运：0001，货运：0010，危化品：0100，其他：1000；违法数据中行为特征转换，违法停车:000001,超速:000010,违法交通信号灯:000100,非法营运:001000,超员:010000,超载:100000等。

6.2.3　标准化数据法

对于不同特征取值范围相差较大的，将特征值通过公式(x-均值)/方差映射到[0,1]范围内，如车辆数据中车龄进行计算得到标准化后数据，违法数据中总违法记分数进行标准化。

6.2.4　分箱处理法

对于连续型特征，转换为类别型的特征。

如对驾驶人年龄做分箱处理，处理后[18-23]、[24-30]、[31-35]、[36-40]、[41-50]、[51-60]、60以上共7个类别；过车数据中近三十天车辆过车天数处理后小于3天、[4-8]、[9-13]、[14-16]、[17-21]、21天以上共6个类别。

6.3　特征筛选

6.3.1　特征重要性排序

使用随机森林算法或决策树算法中的特征重要性计算模块来计算特征重要性，并按照重要性做降序排序，得到特征的重要性集合N。

6.3.2　特征筛选

剔除特征重要性排序最低的10%的特征，得到新的特征集合，用新的特征集合，重复上述过程，直到剩下0.75×N个特征。

7 数据建模

7.1 业务规则类模型

7.1.1 适用场景

检测交通管理业务中假牌车、套牌车、车辆逾期未检验、报废车辆上路行驶等不合规则的场景。

7.1.2 使用数据

7.1.2.1 假牌车识别模型建模时应使用附录 B 中号牌号码、号牌种类、车辆类型和附录 F 中的号牌号码、号牌种类、车辆类型、过车时间、设备编号。

7.1.2.2 套牌车识别模型建模时应使用附录 B 中的号牌号码、号牌种类、车辆类型，附录 F 中的号牌号码、号牌种类、车辆类型、过车时间、设备编号、车道编号和附录 H 中的设备编号、车道编号、点位编号、经度、纬度。

7.1.2.3 车辆逾期未检验识别模型建模时应使用附录 B 中的号牌号码、号牌种类、车辆类型、检验有效期止和附录 F 中的号牌号码、号牌种类、过车时间、设备编号。

7.1.2.4 报废车辆上路行驶识别模型建模时应使用附录 B 中的号牌号码、号牌种类、车辆类型、强制报废期止和附录 F 中的号牌号码、号牌种类、过车时间、设备编号。

7.1.3 设置规则

7.1.3.1 最新过车数据中的车辆在车辆信息中匹配不到数据，则认为此车辆的号牌为假牌。

7.1.3.2 最新过车数据中的车辆在不同的点位同时出现，且两个点位之间距离大于 500 米，认为此车辆为套牌车。

7.1.3.3 最新过车数据中的车辆检验有效期超出了车辆信息中检验有效期止日期，则认为车辆为逾期未检验。

7.1.3.4 最新过车数据中的车辆报废日期超出了车辆信息中强制报废期止日期，则认为此车辆为报废车辆上路行驶。

7.1.4 规则计算

7.1.4.1 在最新过车数据中关联不到车辆信息中的号牌号码、号牌种类、车辆类型数据，则将该辆车定为假牌车，标签值为 1，否则标签值为 0。

7.1.4.2 在最新过车数据中关联到两个点位距离大于 500 米、求和值大于 1 的车辆，则将该组车定为套牌车，标签值为 1，否则为 0。

7.1.4.3 在最新过车数据中关联到车辆检验数据超出了车辆信息中检验有效期止，则将该车定为逾期未检验车，标签值为 1，否则为 0。

7.1.4.4 在最新过车数据中关联到车辆报废数据超出了车辆信息中车辆强制报废期止，则将该车定为强制报废车，标签值为 1，否则为 0。

7.1.5 模型输出

7.1.5.1 筛选模型输出标签值为 1 的数据作为模型识别结果。

7.1.5.2 模型应输出假牌车的号牌号码、号牌种类、过车时间、设备编号信息。

7.1.5.3　模型应输出套牌车的号牌号码、号牌种类、过车时间、设备编号信息。
7.1.5.4　模型应输出逾期未检验车的号牌号码、号牌种类、过车时间、设备编号、车辆检验有效期止信息。
7.1.5.5　模型应输出强制报废车的号牌号码、号牌种类、过车时间、设备编号、强制报废期止信息。

7.2　预测预警类模型

7.2.1　适用场景

驾驶人、机动车和道路有安全风险的场景。

7.2.2　使用数据

7.2.2.1　驾驶人应使用附录 A 中的驾驶人出生日期、性别、身份证明号码、初次领证日期、驾证期限、累计记分、超分日期、准驾车型、驾驶证状态、有效期止、有效期始、补证次数，附录 C 中的违法行为、违法时间、违法记分数、机动车使用性质、号牌号码、号牌种类，附录 D 中的事故发生时间、事故类型、碰撞方式、当场死亡人数、抢救无效死亡人数、重伤人数、轻伤人数、24 小时内死亡人数、3 日内死亡人数、7 日内死亡人数、30 日内死亡人数、机动车数量、非机动车数量、行人数量、事故编号，附录 E 中的事故编号、身份证明号码、驾驶证种类和附录 J 中的身份证号码、是否吸毒人员。
7.2.2.2　机动车应使用附录 C 中的违法行为、违法时间、号牌号码、号牌种类，附录 D 中的事故发生时间、事故类型、碰撞方式、当场死亡人数、抢救无效死亡人数、重伤人数、轻伤人数、24 小时内死亡人数、3 日内死亡人数、7 日内死亡人数、30 日内死亡人数、机动车数量、非机动车数量、行人数量、事故编号，附录 E 中的事故编号、身份证明号码和附录 B 中的号牌号码、号牌种类、身份证明号码、使用性质、强制报废期止、发牌日期、核定载客。
7.2.2.3　道路应使用附录 G 中的道路名称、道路代码、路面结构、行政区划、管理部门、道路类型、道路物理隔离、地形、公路行政等级、路侧防护设施类型、路段代码、路口 ID，附录 D 中的路号、路名、公里数、米数、管理部门、事故类型、事故发生时间、能见度、天气、当场死亡人数、重伤人数、轻伤人数、机动车数量、事故认定原因分类、地形和附录 I 中的管理部门、降雨量、降雪量、平均能见度、平均风速、湿滑系数。

7.2.3　算法选择

7.2.3.1　可选用决策树、随机森林、逻辑回归、K-近邻算法、神经网络、Adaboost、XGBoost、朴素贝叶斯、支持向量机算法、线性分类器算法、梯度提升数算法、高斯混合模型算法等。
7.2.3.2　宜使用 XGBoost 算法。

7.2.4　划分数据集

随机抽取特征集中的 75％数据作为训练集，15％数据作为验证集，10％数据作为测试集。

7.2.5　模型训练

导入XGBoost算法模块，设置为树模型，最小样本权重设置为[0.3,0.8]，损失函数设置为softmax，训练迭代次数设置为50，提前终止迭代次数设置为20，学习率设置为[0.01,0.3]，学习率步长设置为0.05，训练最大深度设置为[5,15]，对训练集进行多轮训练，选取一组训练结果较优的模型参数，使用验证集对模型训练效果做验证。

7.2.6　模型输出

模型输出结果按照驾驶人、机动车和道路均分为重大风险（标签值1）、较大风险（标签值2）、一般风险（标签值3）、低风险（标签值4）四个等级。

7.3 异常检测类模型

7.3.1 适用场景

非现场违法取证设备异常检测、机动车非法营运识别、路口流量激增预警场景。

7.3.2 使用数据

7.3.2.1 非现场违法取证设备异常检测应使用附录 F 中的经过时间、设备编号、车道号、号牌号码、号牌种类、方向，附录 C 中的违法时间、违法代码、管理部门、号牌号码、号牌种类、路口路段代码和附录 H 中的设备编号、车道编号、管理部门、设备类型、点位编号、行政区划。

7.3.2.2 机动车非法营运识别应使用附录 B 中的号牌号码、号牌种类、车辆类型、使用性质、核定载客,附录 F 中的经过时间、设备编号、车道号、号牌号码、号牌种类、方向、号牌颜色，附录 G 中的道路代码、路段代码、道路名称、路口 id、路面名称，附录 H 中的设备编号、点位编号、路口 id，以及通过过车数据构造出的每天平均经过的点位数量，每天过车小时时段数量，平均每天经过的道路条数、一个月内过车天数、平均每天经过不同点位数量、工作时间段过车天数、工作时间段过车天数占过车记录天数比例、工作时间段经过点位数量占全天过车经过点位数量比例、晚上(20:00 后和 07:00 前)时段有过车记录天数。

7.3.2.3 路口流量激增预警应使用附录 G 中的道路名称、道路代码、行政区划、管理部门、道路类型，附录 F 中的过车时间、号牌号牌、号牌种类、设备编号、车道编号和附录 H 中的设备编号、车道编号、路段代码、路口 id、行政区划、管理部门。

7.3.3 算法选择

7.3.3.1 非现场违法取证设备异常检测、路口流量激增预警可选择移动平均法、周期因子法、指数平滑算法、ARIMA、Prophet、RSI、Holt-Winters、RNN、LSTM、seq2seq、DeepAR、WaveNet 等，非现场违法取证设备异常检测模型宜采用 Prophet 算法。

7.3.3.2 机动车非法营运识别模型可选择基于分布的 Z-Score、3sigma、boxplot、Grubbs 假设检验、基于距离的 KNN、基于聚类的 DBSCAN、基于树的 iForest、基于降维的 PCA、AutoEncoder、基于分类的 One-Class SVM，基于密度的 LOF、SOS、COF 等，宜采用基于树的 iForest 算法。

7.3.4 划分数据集

7.3.4.1 时间序列类异常数据，按时间序列顺序选取前 90％的数据作为训练集，后 10％的数据作为验证集。

7.3.4.2 非时间序列类异常数据，随机抽取特征集中的 75％数据作为训练集，15％数据作为验证集，10％数据作为测试集。

7.3.5 模型训练

7.3.5.1 时间序列类导入 Prophet 算法模块，时间序列数据增长趋势设置为 logistic，变化点灵敏度设置为为低，季节性灵敏度设置为高，假期效果灵敏度设置为高，置信度区间设置为[0.8,0.85]，步长设置为 0.01，季节性周期设置为[月,季度]，变化点数量设置为[25,35]，步长设置为 1，假期日期加入中国法定节假日。超参数设置完成后，对模型迭代训练，训练迭代次数设置为 50，选取一组训练结果较优的模型参数，使用验证集对模型训练效果做验证。

7.3.5.2 非时间序列类导入 iForest 算法模块，基本估算器 n_estimators 数量设置为[100,200]，最大样本数量 max_samples 设置为 auto，最大特征数量设置为[5,10]，训练迭代次数设置为 50，选取一组训练结果较优的模型参数，使用验证集对模型训练效果做验证。

7.3.6 模型输出

7.3.6.1 时间序列类输出的值大于预测值的 30%为异常，非时间序列类输出标签值为-1 的为异常。
7.3.6.2 模型应输出异常的非现场违法取证设备的设备编号、异常检测值、发生时间信息。
7.3.6.3 模型应输出路口流量激增的路段代码、路口 id、车道编号、监测时间、预测流量信息。
7.3.6.4 模型应输出机动车非法营运的车辆号牌、号牌种类及异常值标签信息。

8 模型评估

8.1 评估指标

8.1.1 预测预警类模型

8.1.1.1 可选用混淆矩阵、精确率、准确率、召回率、F1 值、AUC 值、ROC 曲线、PR 曲线。
8.1.1.2 宜使用 F1 值作为评估指标。
8.1.1.3 F1 值计算方法。

a) 用验证集数据中模型预测为真正正确样本(TP)的个数除以所有预测为正确样本个数(TP+FP)；
b) 模型预测为真正正确样本（TP）的个数，除以所有的实际为正样本个数（TP+FN）；
c) 根据查准率 P 和查全率 R，等到 F1 值=2×P×R/(P+R)。

8.1.2 异常检测类模型

8.1.2.1 时间序列类可选用均方根误差(RMSE)、平均绝对偏差(MAE)、偏差(BIAS)、相关系数(CORR)和准确率（ACCURATE），宜采用均方根误差（RMSE）作为评估指标。
8.1.2.2 非时间序列类可选用混淆矩阵、精确率、准确率、召回率、F1 值、AUC 值、ROC 曲线、PR 曲线，宜使用 AUC 值作为评估指标，通过计算 ROC 曲线下的面积，得到 AUC 值，AUC 取值范围为[0,1]。

8.2 评估方法

8.2.1 业务规则类模型

8.2.1.1 业务规则类模型评估可采用数据源的验证、比对实际业务流程和规则冲突检测方法。
8.2.1.2 数据源的验证应对业务数据的准确性进行验证。
8.2.1.3 比对实际业务流程应通过与实际的业务规则进行比对和核对，判断业务规则类模型的准确性和适用性。
8.2.1.4 规则冲突检测应对业务规则类模型中是否存在冲突或重复进行判断，并调整或修正。

8.2.2 预测预警类模型

8.2.2.1 预测预警类模型评估方法可采用交叉验证法、留出法、自助法。
8.2.2.2 交叉验证发应使用 K 折交叉验证，将所有训练数据集分成 K 个大小相当的子样本，取其中一个子样本作为验证集，其余 K-1 个作为训练集，最后对 K 次建模的结果进行综合评价，验证每轮训练的模型预测效果，以及对应的参数值设置，选择最优的预测效果对应的参数值，作为模型最优参数，K 值可以选择[5,10]之间的数值。