

基于XDR大数据分析和AI技术的 定轨道路用户感知识别技术

User Perception Recognition of Fixed Track Road Based
on XDR Big Data Analysis and AI Technology

潘 晖,齐咏嘉,杭旭峰,姚赛彬,黄久成(中国联通上海分公司,上海 200080)

Pan Hui, Qi Yongjia, Hang Xufeng, Yao Saibin, Huang Jiucheng (China Unicom Shanghai Branch, Shanghai 200080, China)

摘 要:

随着网络结构、业务形态、用户模型的不断变化以及大数据、云计算等技术的不断演进,网络运维成本压力逐年上升,依赖于专业测试软件进行道路测试的传统优化模式已经难以满足未来发展的需要,网络优化亟待智能转型。基于定轨道路用户的真实海量数据,通过聚类算法模型对上海全路段用户感知问题进行定位分析研究。研究表明,大数据分析和AI算法的定位精度和优化效率远大于传统路测。

关键词:

智能转型;聚类算法;大数据分析;用户感知

doi: 10.12045/j.issn.1007-3043.2021.03.017

文章编号: 1007-3043(2021)03-0077-07

中图分类号: TP391

文献标识码: A

开放科学(资源服务)标识码(OSID):



Abstract:

With the continuous change of network structure, business form, user model and the continuous evolution of big data, cloud computing and other technologies, the pressure of network operation and maintenance cost is increasing year by year. The traditional optimization mode relying on professional test software for road test is difficult to meet the needs of future development, and network optimization needs intelligent transformation. Based on the real mass data of rail users, it analyzes and studies the user perception problem location of the whole road section in Shanghai by clustering algorithm model. The results show that the positioning accuracy and optimization efficiency of big data analysis and AI algorithm are much higher than those of traditional road test.

Keywords:

Intelligent transformation; Clustering algorithm; Big data analysis; User perception

引用格式: 潘晖,齐咏嘉,杭旭峰,等. 基于XDR大数据分析和AI技术的定轨道路用户感知识别技术[J]. 邮电设计技术, 2021(3): 77-83.

0 前言

传统的道路优化采用DT路测分析加KPI指标监控的方式。DT路测分析需要测试工程师使用专业的测试软件和测试设备模拟用户使用情况沿一定道路进行现场实地测试,这种优化方式成本高,路测数据样本量小,具有时间、地点的随机性,无法模拟真实用户感知。KPI指标监控是先筛选出定轨道路沿线主控小区然后进行大量KPI指标性能统计,这种方式会引

入大量非定轨道路用户,同时也需要投入大量人力物力,准确性和效率都大打折扣。而高架、高铁、高速、地铁等快速道路由于用户行驶路线相对固定且行驶速度相对较快,对于用户感知的要求更高。此外2G/3G/4G的多网融合,以及5G网络的垂直应用等多方面都对网络优化的响应支撑能力提出新的挑战。

为解决传统路测优化方式固定、周期性长、工作量大、耗时等问题,网络优化的智能转型刻不容缓。以上海高架道路为例,根据最新统计上海高架日均车流量已经达到201万,面对如此大的样本数量,网络优化工作能否引入大数据分析和AI智能识别的方式,精

收稿日期: 2021-01-15

准定位出定轨道路用户呢? 本文基于真实定轨道路用户的海量业务分析, 将用户感知情况自动关联至小区, 为网络优化提供充分数据依据, 将优化模式由传统的线下模式转型为线上模式, 从而大幅提高网络优化效率, 有效降低优化成本。

1 定轨道路感知识别技术介绍

本文运用用户感知识别算法主要结合前期全上海定轨道路传统DT路测样本指纹库数据, 通过自动采集上海联通移网用户的XDR、MR、CDR等数据与指纹库进行运动特征判定, 并运用聚类分析方式实现定轨道路建模和用户识别, 通过Python进行环境搭建, 嵌入PostgreSQL数据库应用以及KNIME和Tableau的大数据挖掘和可视化工具, 最终实现定轨道路用户感知识别。基于大数据分析和AI技术识别方式相对传统的道路测试方式数据采集效率更高、成本更低、数据更丰富, 也能更真实地反映现网用户的网络感知。

1.1 AI算法选择

定轨道路用户感知识别是一种典型的聚类分析算法应用场景。聚类就是按照某个特定标准(如距离准则)把一个数据集分割成不同的类或簇, 使得同一个簇内数据对象的相似性尽可能大, 同时不在同一个簇中的数据对象的差异性也尽可能的大。即聚类后同一类的数据尽可能聚集到一起, 不同类数据尽量分离。

聚类技术通常又被称为无监督学习, 因为与监督学习不同, 在聚类中是没有那些表示数据类别的分类或分组信息的。

a) 聚类(Clustering): 简单地讲就是把相似的东西分到一组, 聚类并不关心某一类是什么, 只需要把相似的东西聚到一起。因此, 一个聚类算法通常只需要知道如何计算相似度就可以开始工作了, 因此聚类通常并不需要使用训练数据进行学习, 这在机器学习中被称作无监督学习(unsupervised learning)。

b) 分类(Classification): 对于一个 classifier, 通常需要告诉它“这个东西被分为某某类”这样一些例子, 理想情况下, 一个 classifier 会在它得到的训练集中进行“学习”, 从而具备对未知数据进行分类的能力, 这种提供训练数据的过程通常叫做监督学习(supervised learning)。

定轨道路用户识别由于没有足够多的样本数量作为训练集, 是一种典型的无监督学习的应用场景,

非常适用聚类分析算法。

1.2 定轨道路用户数据建模

选定算法以后, 需要进行定轨道路用户数据建模, 具体实现步骤如图1所示。

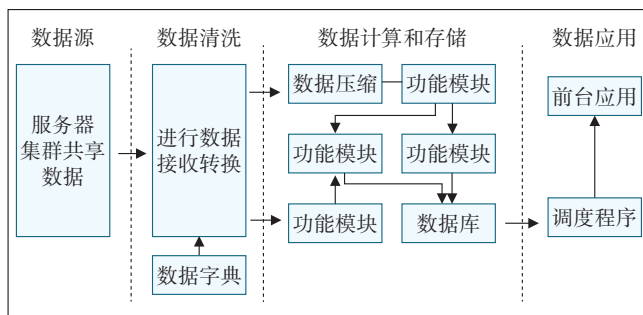


图1 数据建模

a) 对每个用户的业务详单进行排序, 然后根据运动时序及站点更新进行运动特征建模。

b) 对建模的数据进行数据清洗和一致性检查。

c) 计算出每个用户的运动模型后与特征指纹库进行比对聚类出定轨道路用户, 最终得出聚类模式。

数据源方面本文引入了XDR数据, 数据量由以前的每日2亿的语音CDR数据提升至每日80亿的XDR数据, 判断运动特征的方法也由传统KPI的小时粒度切片方式升级为业务遍历方式。最后通过从全量XDR数据匹配重点道路工参表, 运用聚类算法计算出每个用户每天的运动轨迹, 判断用户是否为运动状态。

1.3 用户运动数据特征判定

基于移动性原理, 用户在定轨道路移动的过程中会从一个位置区移动到另一个位置区, 位置更新的数据将被记录为网络侧大数据。如果网络侧用户在短时间内发生多个位置区更新, 则将其定义为运动特征用户, 同时将位置更新的时间间隔与定轨道路运行时长匹配, 将定轨道路用户从大网用户中分离出来并进行聚类分析。当用户每次占用的基站发生改变, 计算出当前与上次转换是否超过15 min, 如果没有超过则运动继续, 否则此次运动结束, 基于此算法来获取大量的用户运动行程碎片。

1.4 数据清洗和标准化

得到大量的用户运动碎片之后, 首先对用户行程碎片进行时序排列, 记录两端用户碎片时间。将第1个运动碎片占用的最后一个基站和下一个运动碎片占用的第一个基站进行经纬度匹配。通过计算基站

经纬度站距与运动碎片相邻时间差的比值得到用户运动速度,如果运动速度大于 30 km/h,则认为用户在 2 个运动碎片之间依然保持运动状态,两端用户碎片

可以合并,最终形成完整的用户运动轨迹和运动时间(见图 2)。

1.5 特征指纹库搭建

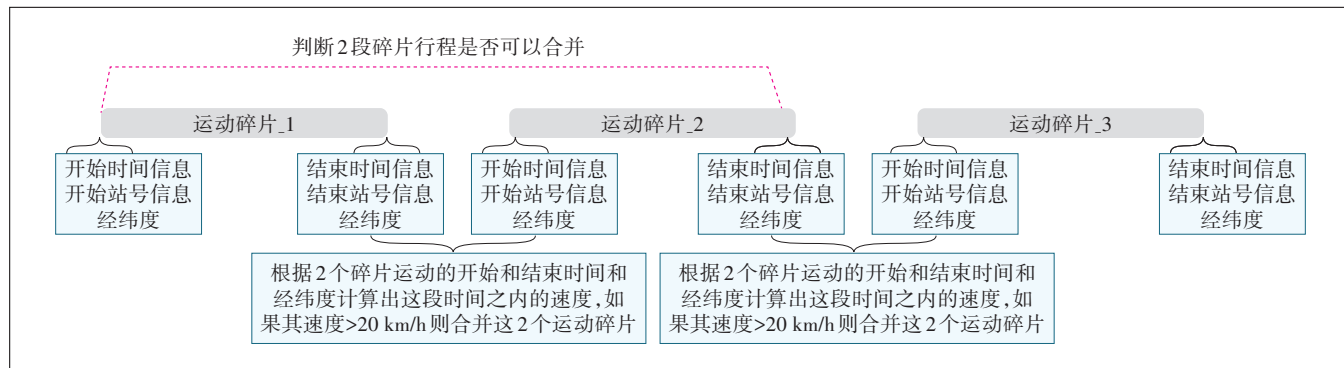


图2 运动碎片合并

1.5.1 区块化切割

如图 3 所示,根据定轨道路和站点路段匝道分布,将上海高速、高架、高铁、地铁场景切割为 2 263 个区块化路段,其中 14 条高架切分成 174 条具体区块路段。



图3 高架路段分割图

1.5.2 指纹库识别

以 DT 数据作为指纹库,将区块路段按照实际覆盖距离进行道路路段细分,如图 4 所示。

1.5.3 定轨道路用户聚类

将实际业务详单与 DT 指纹库进行匹配,并通过用户运动轨迹匹配定轨道路轨迹,如果用户连续占用该道路的 3 个路段 ID,则将用户聚类为该定轨道路路段用户,例如延安高架道路一共有 10 个不同的路段 ID(每个路段 ID 对应多个站点),如果一个运动行程中

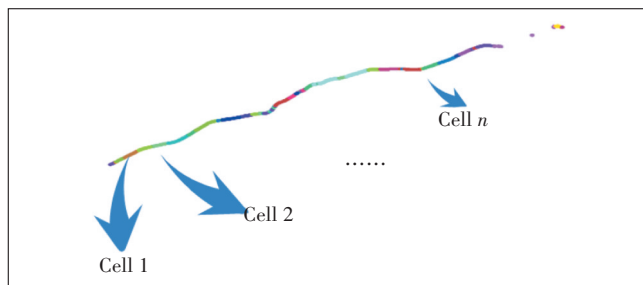


图4 DT 指纹库

占用大于 3 个路段 ID 就判断他为延安高架用户,并识别出其在延安高架上的开始、结束时间和开始、结束地点。每个运动行程会和所有重点道路做匹配,满足匹配要求即为该重点道路用户,如图 5 所示。

定轨道路用户需要至少占用 3 个连续路段,如果少于 3 个路段就进行定轨道路识别,会引入大量非道路用户,导致自动识别准确度下降。例如一个非定轨道路用户在路段 A 与路段 B 之间通话,他同时占用路段 A 与路段 B 的主控基站,该用户也会被聚类为定轨道路用户,导致识别准确度下降。

1.6 模型应用

本文的用户模型改变了以往传统时间切片方式,由业务遍历方式进行用户画像识别(见图 6)。该技术以用户感知驱动诊断网络问题关联 MR 数据和信令数据,端到端分析定轨道路网络感知。

该技术通过手机上报的 GNSS 经纬度信息以及相关优化参数,辅助网络精准分析、优化,结合用户 CDR 话单预测用户轨迹。对于室外用户,利用电子地图索引和 CDR 样本的轨迹预测出用户实际经过的道路,并确认用户发生问题的精确位置,同时对模型识别用户

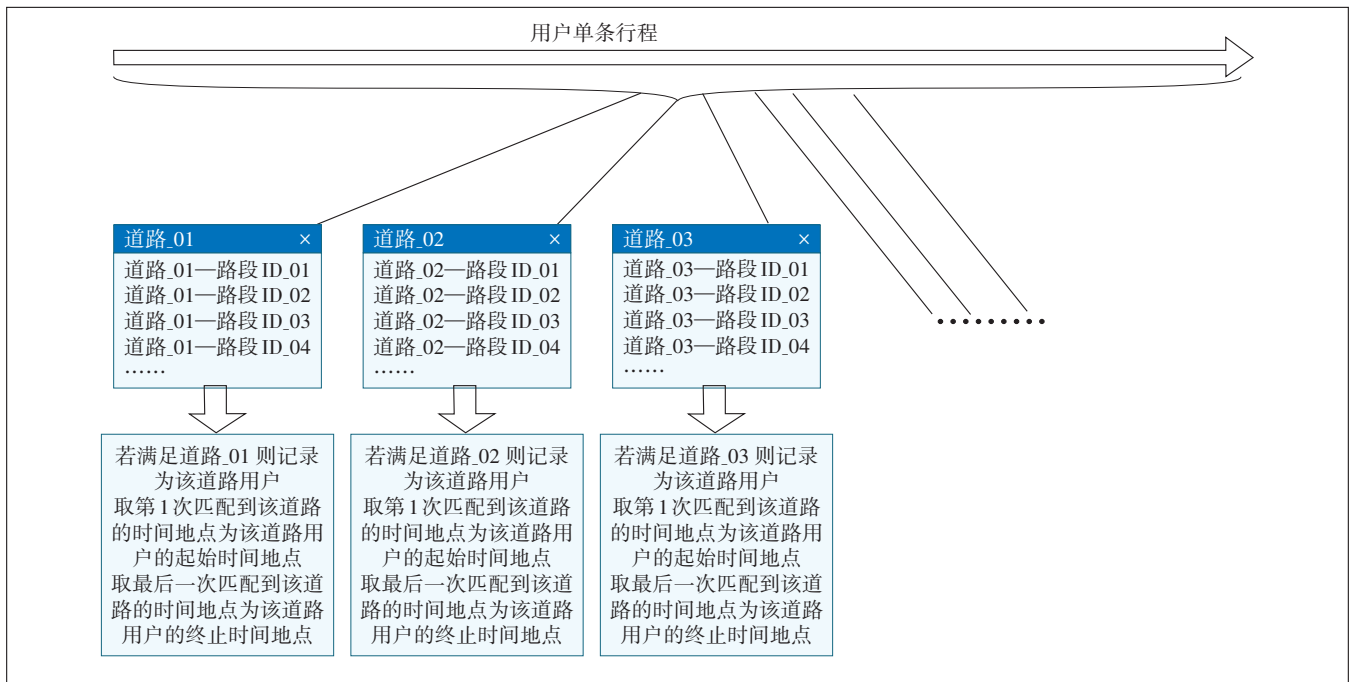


图5 定轨道路匹配

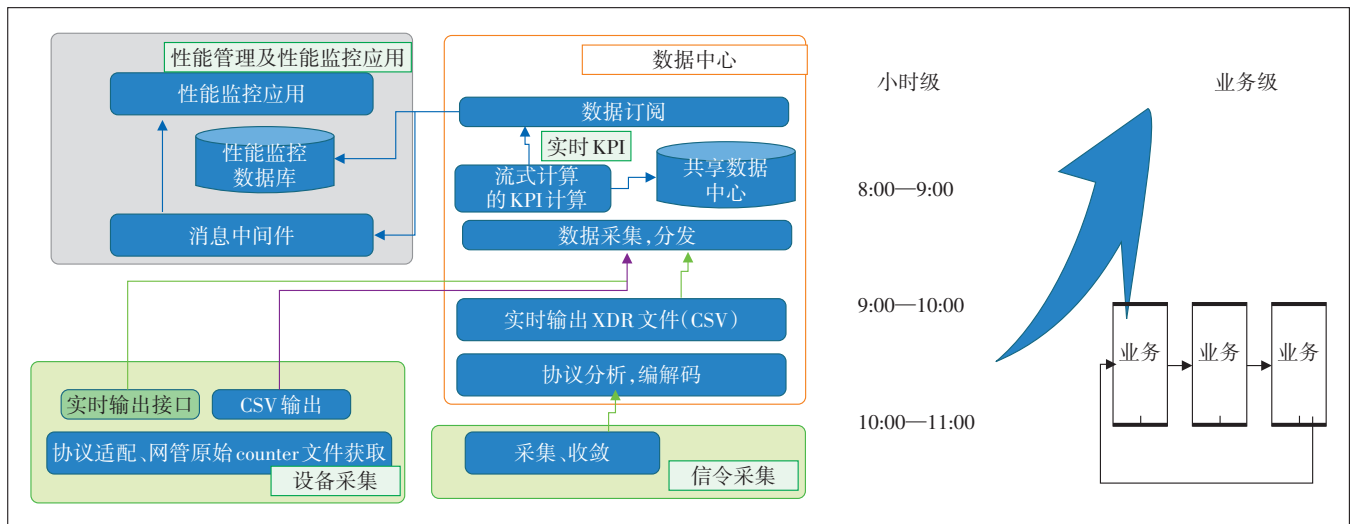


图6 业务级遍历模式

进行空间定位,并投射到GIS中。该技术通过对用户问题点进行汇聚,发现问题路段或用户投诉风险路段并进行预警。

通过用户手机号码,实现定轨道路用户与业务质量、终端数据、3G MR、XDR数据关联,精准定位网络问题(见图7)。

1.7 大数据处理流程图

基于XDR的用户行为模型分析,辅助匹配道路特征指纹库及用户感知话单数据,实现定轨道路MR覆盖和语音感知智能评估。

由于全网用户每天的XDR数据达到10亿级,需搭建5台服务器进行SEQ数据处理,每台机器共有100个分区表,也就是每台机器都要调用100次存储过程。由于数据量庞大,可通过定时任务的方式每天凌晨进行数据聚类,每天数据处理时间达到6h以上。最终输出道路级、用户级分析结果。各个环节处理流程如图8所示。

2 可视化展示

本文介绍的基于XDR大数据分析和AI技术的定

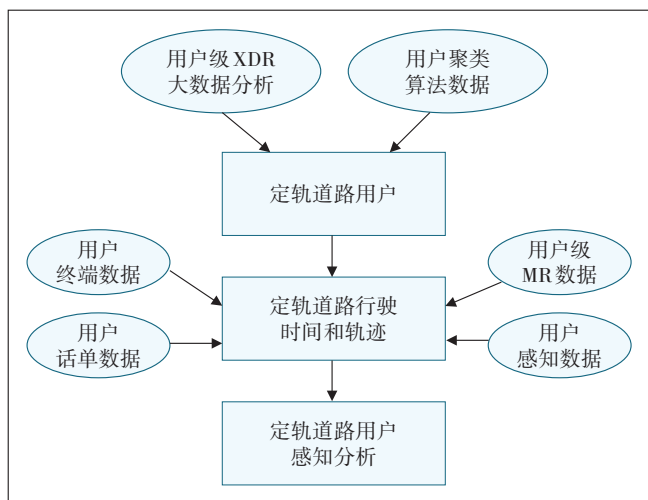


图7 多维度数据关联

轨道路用户感知识别技术通过对终端上报的XDR信息进行精准定位,结合电子道路与采样点拟合技术,突破了传统的单一终端主干道测试的模式,实现全民测试、全量测试。该技术通过平台化自动分析处理,快速汇聚问题道路,并提供友好界面呈现。目前基于XDR大数据分析和AI技术的定轨道路用户感知识别技术已投入上海联通网络优化中心实际生产中,通过可视化大屏可实现实时KPI监控和单用户感知定位分析,可视化监控系统如图9所示。

用户级感知情况可以在tableau大屏中通过SQL语句对单用户行程进行查询,得到如表1所示的结果。

3 定轨道路感知识别技术的价值

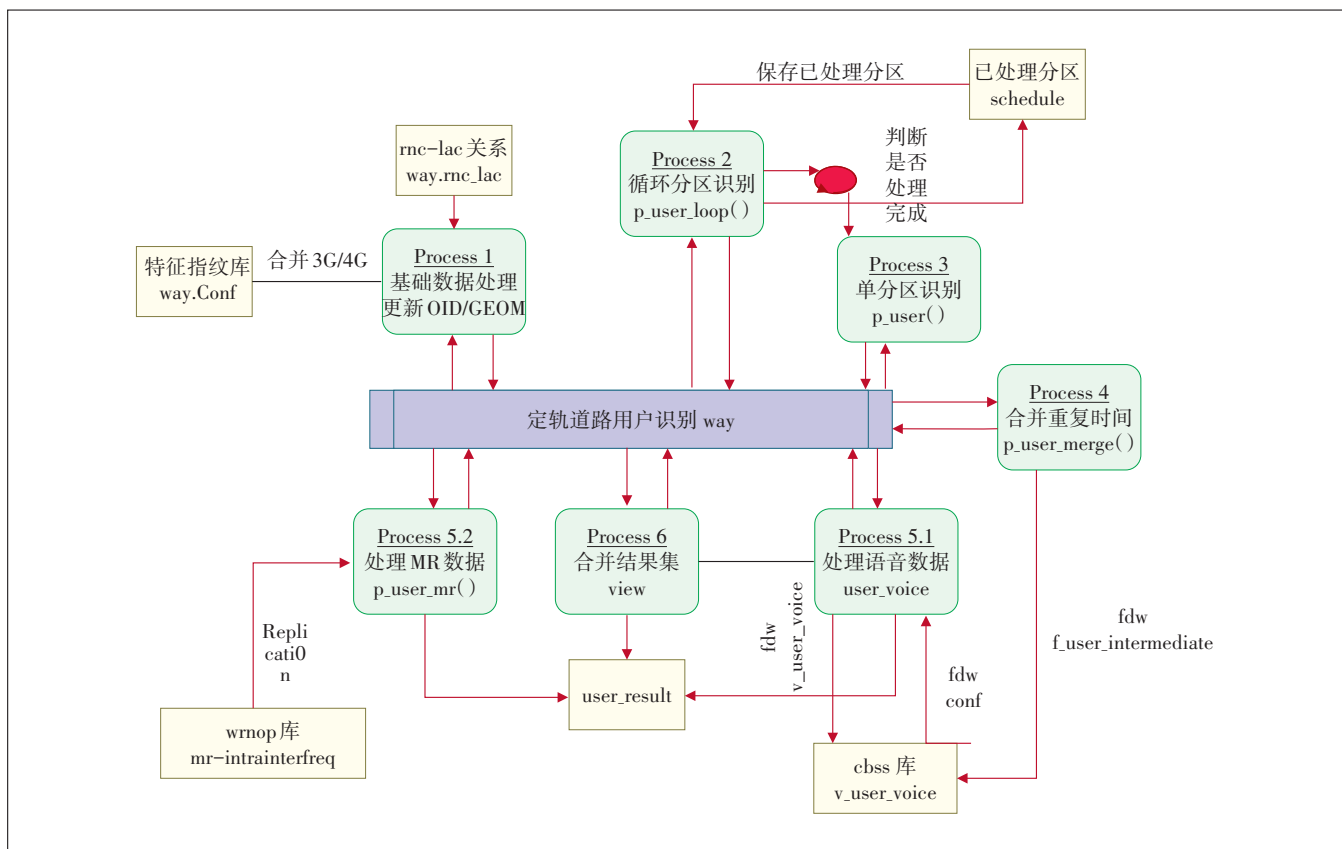


图8 数据建模执行流程图

基于XDR大数据分析和AI技术的定轨道路用户感知识别技术已纳入网络优化实际生产中,为优化工程师助力提效。

3.1 应用价值

通过此算法,系统可每日自动输出高掉话路段、质差路段,纳入日常优化管控表,形成从智能识别到

自动诊断再到优化闭环的问题点管控机制。值得一提的是,自动识别出的问题点中有些是日常优化过程中未发现的,这帮助优化工程师发现了网络的隐性问题。如表2所示,目前经过算法的调优和改进,系统可日均智能识别定轨道路用户30 000个以上。

3.2 推广价值



图9 可视化监控系统

表1 单用户运动轨迹查询

用户号码	定轨道路	开始时间	结束时间	行驶路段数
186xxxxxxxx	内环高架	2020-04-23 08:06:32	2020-04-23 08:26:45	6
	中环高架	2020-04-23 08:34:41	2020-04-23 08:41:06	4
	虹梅南路高架	2020-04-23 08:55:30	2020-04-23 09:11:12	3
	中环高架	2020-04-23 15:43:52	2020-04-23 15:50:17	4
	内环高架	2020-04-23 15:54:40	2020-04-23 16:02:07	5

表2 各路段识别用户数

定轨道路	识别用户数	语音通话次数	XDR掉话次数	XDR掉话率/%
嘉闵高架	1 300	2 913	7	0.25
罗山路高架	628	1 586	3	0.21
五洲高架	355	918	2	0.21
沪闵高架	710	1 725	3	0.20
中环高架	6 868	16 120	30	0.19
北翟高架	521	1 289	2	0.16
内环高架	6 599	15 739	22	0.14
延安高架	2 969	7 099	10	0.14
南北高架	3 197	7 590	10	0.13
逸仙高架	1 481	3 785	5	0.12
外环高架	4 992	12 261	13	0.11
华夏高架	723	1 922	2	0.10
虹梅南路高架	513	1 222	1	0.05
迪士尼高架	65	130	0	0.00
总计	30 856	74 175	109	0.15

自XDR大数据分析和AI技术的定轨道路用户感

知识识别技术纳入网络优化生产以来,上海联通实现了80%的路测优化自动执行,且规范了优化方法和手段,大幅提升了网络优化工作效率。质差及高掉话等用户感知隐性问题的发现及时率从30%提高到95%,网络优化效率提升90%,日常优化问题点覆盖率达到90%以上;工作自动执行可完成率达到70%以上,节约了运维成本,提高了优化效率。目前已累计识别用户感知类优化问题点220个,表3列出了部分问题路段。

后续通过指纹库的扩建精准化定位,可以从定轨道路场景识别推广至省一级以上干线道路场景。同时随着关联的数据越来越丰富,识别的用户业务也越来越多样,后续将引入KQI数据和VoLTE话单甚至5G业务等数据,从语音感知端到端分析,扩展到VoLTE、视频、游戏等多业务感知识别。

3.3 经济价值和社会效益

本文提出的定轨道路感知识别技术带来的经济

表3 高掉话路段示例

高架名	路段名	语音识别出的用户数	总通话次数	掉话次数	掉话比/%	是否路测问题点
外环高架	沪青平公路—吴中路	115	116	4	3.45	是
嘉闵高架	扬虹路高架—申兰路	126	135	4	2.96	是
嘉闵高架	真南路—曹安公路	30	39	1	2.56	是
外环高架	济阳路—上南路	37	40	1	2.50	是
中环高架	吴中路匝道—仙霞路匝道	99	104	2	1.92	是
华夏高架	华洲路—浦东机场连道晨阳路匝道	97	111	2	1.80	否
外环高架	丰翔路—真南路	59	62	1	1.61	否
内环高架	歇浦路—河间路	377	441	7	1.59	是
逸仙高架	殷高路—长江南路	57	68	1	1.47	是
南北高架	威海路—淮海中路	87	90	1	1.11	是
南北高架	场中路—中环交汇处	416	487	5	1.03	是

效益如下。

a) 统一采集存储和计算,提高资源利用率,减少重复建设。

b) 减少测试工作的人员和设备的支出,释放优化分析人员成本。

c) 本文提出的算法是自主研发,锻炼了网络优化工程师的自主开发能力,节省了工程建设费用。

d) 通过定轨道路用户的自动识别降低路测成本,全网定轨道路评估能力从每月1次提升到每天1次,全年节省道路测试1 000次,节省费用测试560万元(全网单次路测成本约为15 000 km×30 km/元=45万元)。

本文提出的定轨道路感知识别技术带来的社会效益如下:

基于XDR大数据分析和AI技术的定轨道路自动感知识别,实现了语音感知监控可视化,深化了五心服务,体现了中国联通的匠心网络精神,在业界取得很好反响。经过上海联通的实际生产推广,对其他省分、电信友商提供了很好的参考借鉴。网络质量的改善,降低了用户投诉率,提升了用户满意度,对和谐社会的建设也起到积极作用。在2018年上海第一届世界进口博览会上,本文提出的基于XDR大数据分析和AI技术的定轨道路自动感知识别可视化大屏,部署在

市通信管理局、市进博会保障中心,其对网优指标的可视化监控保障支撑,以及对进博会的安全保障都起到了积极辅助支撑作用。

4 结束语

在网络优化工作中可通过大数据技术进行多维度数据的融合来打破数据壁垒,实现更多的数据碰撞,衍生百花齐放的应用。网络优化工作能够依托现有大数据平台和AI技术,将网络中海量的MR、MDT、XDR、CDR等数据关联融合并实现共享,将用户感知问题从小区级衍生到用户级。网络优化工作需要从用户感知出发,从常规业务、用户、覆盖、容量、质量及3G/4G等多维度进行数据深度关联处理,通过大数据处理和AI等技术的运用,支撑面向规划、面向网络、面向客户、面向市场的相关工作,这才是增效降本的有效方法。

参考文献:

- [1] 李伦波,马广富. 基于PNN的退化交通标志图像的识别算法研究[J]. 电子与信息学报,2008,30(7):1703-1707.
- [2] 陈德军,罗金成,张兵. 基于改进的K-means聚类算法的分类评价方法[J]. 武汉理工大学学报(信息与管理工程版),2011(1):32-35.
- [3] 李伦波,马广富. 自然场景下交通标志检测与分类算法研究[J]. 哈尔滨工业大学学报,2009(11):29-33.
- [4] 赵锋. 数据挖掘技术及其在超市CRM中的应用[D]. 西安:西北工业大学,2007.
- [5] 陈治. 基于GIS的物流配送系统的研究与应用[D]. 杭州:浙江大学,2006.
- [6] 石晶,胡明,戴国忠. 基于小世界模型的中文文本主题分析[J]. 中文信息学报,2007,21(3):69-75.
- [7] 李世峰,黄磊,刘昌平. 几种聚类方法的比较[C]// 全国汉字识别学术会议. 2002.
- [8] 余琳,邓玲. 常见聚类算法的比较以及DSS系统中的应用[C]// 湖北省通信学会,武汉通信学会2009年学术年会. 2009.
- [9] 刘位龙. 面向不确定性数据的聚类算法研究[D]. 济南:山东师范大学,2011.
- [10] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

作者简介:

潘晖,工程师,学士,主要从事WCDMA/LTE无线网络优化相关工作;齐咏嘉,工程师,硕士,主要从事WCDMA/LTE无线网络优化相关工作;杭旭峰,工程师,学士,主要从事WCDMA/LTE无线网络优化相关工作;姚赛彬,工程师,硕士,主要从事WCDMA/LTE无线网络优化相关工作;黄久成,工程师,硕士,主要从事WCDMA/LTE无线网络优化相关工作。