基于人工智能的基站网络流量预测模型设计

徐运武

（广东松山职业技术学院，广东韶关 512126）

【摘要】通过人工智能算法深度挖掘网管数据、用户级数据、经分数据等各数据之间的关系，找到与流量数据相关的向量或参数作为模型训练的特征构建基站预测模型，提前预知基站即将突发的高负荷，提前扩容，提高用户感知，实现基站流量预测，能较好地解决基站当前面临的容量问题。

【关键词】人工智能；基站；流量预测

【doi:10.3969/j.issn.2095-7661.2022.04.002】

【中图分类号】TN929.5　　【文献标识码】A　　【文章编号】2095-7661（2022）04-0004-05

Design of Base Station Network Traffic Prediction Model Based on Artificial Intelligence

XU Yun-wu

(Guangdong Songshan Polytechnic College, Shaoguan, Guangdong, China 512126)

Abstract: Through the artificial intelligence algorithm, the relationship between network management data, user-level data, economic analysis data and other data is deeply mined, and the vector or parameter related to the traffic data is found as the characteristics of the model training to build the base station prediction model, which can predict the sudden high load of the base station in advance, expand the capacity in advance, improve the user perception, realize the base station traffic prediction and better solve the current capacity problem of the base station.

Keywords: artificial intelligence; base station; traffic prediction

随着5G网络覆盖提升及不限量套餐的推出，网络容量成为影响用户感知的主要因素。5G网络容量的分析、预测和扩容方案制定是网优工作的重点。为提升工作效率，通过机器学习来形成更为精准的容量预测和扩容方案将是未来流量预测的主要方向。

通过分析当前基站容量面临的问题，根据现有数据，如网管数据、用户级数据、经分数据等，采取人工智能算法深度挖掘各数据之间关系对基站未来的流量进行预测，对即将突发高负荷基站进行提前预知，提前扩容，提前保障，进一步提高用户感知，实现基站流量预测。

1 现有基站容量面临的问题

基站容量是有限的，随着用户流量的不断增加，需要对基站容量进行扩容或者负载均衡来缓解容量压力。对于运营商而言，需要知道哪些基站的负荷将要达到上限，提前预警并做扩容解决方案，传统的方法是利用统计学思想做线性统计来预测未来的基站流量，但是流量往往不是线性增长的，而是非线性增长的，而且会有突发事件导致流量的激增，预测的难度还是比较大的。人工智能的发展使深度学习技术日趋成熟，可针对非线性数据来构造模型并预测，本设计将采用深度学习的算法模型来预测基站容量。

[收稿日期] 2022-06-16

[作者简介] 徐运武（1979—），男，湖南邵阳人，广东松山职业技术学院高级实验师，研究方向：电子与通信工程。

[基金项目] 2020年广东省教育厅重点领域专项“基于人工智能的网络优化研究与应用”（项目编号：2020ZDZX3113）；2021年广东省教育厅普通高校特色创新项目“新型微雷达在智能停车场中的研究与应用”（项目编号：2021KTSCX225）；2018年韶关市科学技术局项目“无线信号覆盖最优设计研究”（项目编号：2018sn062）。

2 基于人工智能的基站流量预测模型

2.1 实现框图

采用“Hadoop+Hive+Python”的技术架构，如图1所示，既能保证大吞吐量的底层数据处理，也能保障灵活的流程预测模型建立和预测的能力，使整个系统高效稳定运行。

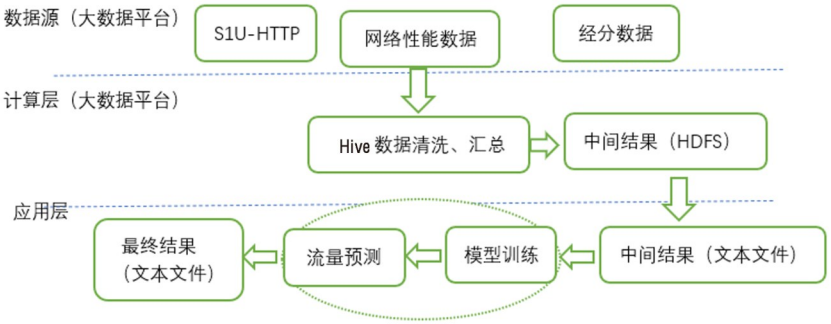

图1　基站流量预测架构图

Hive是Hadoop家族的一个数据仓库工具，是一个SQL语句的解析框架。它封装了一些比较常见的MapReduce任务，并能够像执行SQL一样对存储在HDFS中的表进行操作。Hive有两种类型的表，外表与内表。创建外部表是用来存储数据路径，不对数据的位置进行任何更改[1]，创建内表是用于将数据移动到数据仓库所指向的路径。外部表相对于内部表，可以管理删除元数据，而不删除数据，所以外部表安全等级比内部表更高，更容易对源数据进行共享，数据组织灵活性方面更优秀。

2.2 实现步骤

基站流量预测首先要获取数据，其次是做特征工程，然后是模型训练，最后可根据模型实现流量预测，具体应用功能架构如图2所示。

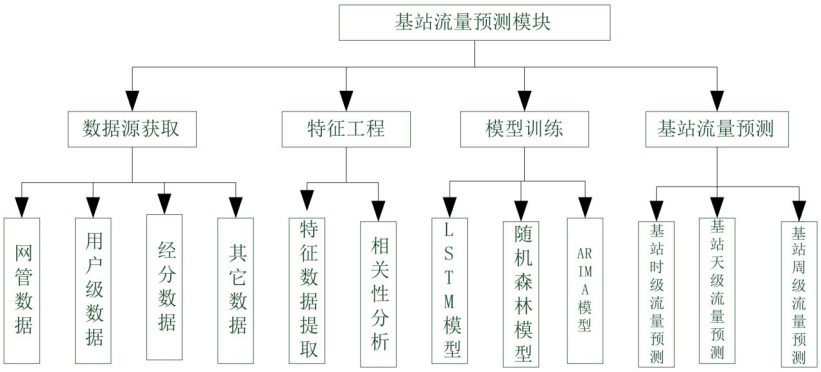

图2　基站流量预测功能架构图

Hive实际上是一个SQL语句的解析框架，提供了一种类SQL语言HQL以便于用户使用，将SQL语言转变成MapReduce进程进行处理。

2.2.1 数据源获取

如表1所示，基站流量预测模型获取网管数据、用户级数据、经分数据及其它数据。网管数据：网管数据主要是监控网络设备的运行状况及查看网络参数的配置是否合理[2]，保障移动网络的正常运行，包含小区ID、Rank 2占比、下行平均误块率、无线资源不足导致的E-RAB连接建立失败率、下行每时隙调度业务PRB数、上行每时隙调度业务PRB数等，采用网管数据可以很直观地看到基站流量变化对网络设备的负载影响。用户级数据：指的是运营商的S1_U-HTTP接口数据，包括用户网络标识、IMSI、CELLID、UL Data、DL Data、UL IP Packet、DL IP Packet、App Type、App Sub-type、App Content等参数，移动用户的上网数据都会通过用户级数据体现出来，采用用户级数据可分析用户行为对基站流量的影响。经分数据：经分数据是用户的套餐信息，可通过此数据分析不同的套餐策略是否对基站流量产生影响。

表1　基站流量预测数据源

数据源名称	数据源类型	数据源格式	存储周期需求	需求字段数
用户级数据	HDFS	.txt/.csv	3天	11
网管数据	HDFS	.txt/.csv	365天	28
经分数据	HDFS	.txt/.csv	全量	13

2.2.2 特征工程

根据相关性分析找到与流量数据相关的向量或参数作为模型训练的特征。这里采用皮尔逊算法选取出与流量相关性最强的9个特征，同时采用随机森林算法模型可得出同样的9个特征，后续根

据这9个特征做流量预测模型：上行PRB平均利用率、下行PRB平均利用率、无线利用率、RRC连接平均数、RRC连接最大数、有效RRC连接最大数、有效RRC连接平均数、上行流量、下行流量。

从原始数据中进行提取特征的方法有多种，这里主要采用相关性分析的方法来对各项数据进行特征的提取。数据项的相关性分析方法用于分辨特征项与分析目标间的相关性，相关性越强则意味着特征项与分析目标项的变化趋势越同步[3]。特征项若与空口上下行流量项的相关性越强，则是该特征可能保留，而对相关性比较弱的特征则可以直接去掉，因为这部分数据若选做特征参与到后面的模型训练，不但使得内存损耗增大，还增加训练集的噪声，影响模型的整体训练效果，并无益处。这里人为定义相关性强弱的判定标准为：相关系数为0.8—1表示相关性最强；相关系数为0.6—0.8表示强相关；相关系数为0.4—0.6表示中等程度相关；相关系数为0.2—0.4表示弱相关；相关系数为0.0—0.2表示极弱相关或无相关。在通过特征项与分析目标项的相关系数选择出特征后，再进行各特征间的协方差矩阵计算，对于两个特征项相关性较高的进行选择性保留，以免产生多重共线性，这样完成特征的最终筛选工作。依据此标准，选择与上下行空口流量强相关的特征项，一共为9项。针对本数据场景，特征工程还需要对数据进行标准化处理。标准化的目的是将数据压缩到一定的标准范围内，防止数据太过于离散，可加快梯度下降求最优解的速度，提高预测精度。

2.2.3 模型训练

利用随机森林模型做特征的选取，利用LSTM模型和ARIMA模型做流量预测的参数训练。评估模型预测效果好坏的标准是SMAPE，一般取值在1以下代表预测效果较好，1以上代表预测存在误差较大，可参考这个标准值去反复训练模型。

2.3 基站流量预测的算法

算法模型部署架构如图3所示，根据容量预测数据源要求的字段格式提供http数据、网管数据、经分数据，数据格式统一整理成CSV格式导入Hive，在Hive里对各数据源做关联处理并汇总成小区小时流量表、小区忙时流量表、小区天流量表、小区周流量表，汇总数据保存成CSV格式传送给分析服务器的流量训练模块，对每个小区单独训练生成流量预测模型，通过流量预测模型对小区未来天、周、月流量进行预测并保存成CSV格式。流量模型训练调优完成后固化训练模型，当新数据到来后可直接进入流量预测模型进行预测，为提升模型的泛化能力，每过一段时间对模型准确性再做验证，当模型不再适合当前的流量预测时，重新对模型进行训练，流量预测采用LSTM神经网络和ARIMA模型，可根据小区特性来选择。

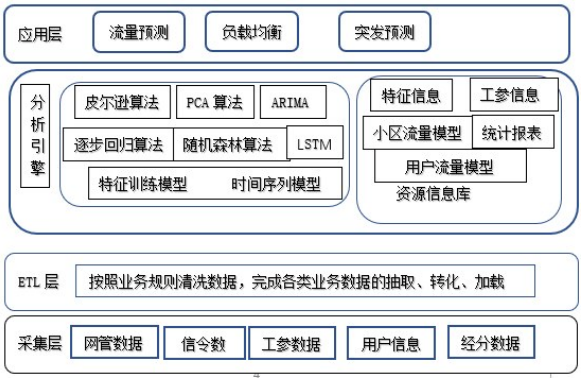

图3 算法模型部署架构图

2.3.1 随机森林算法模型

随机森林是一种灵活且比较常用的集成学习算法，该模型即使使用初始参数，在很多时候也可以得到很好的结果。它也是机器学习中比较常用的算法之一，因为它结构简单却又能使多个弱算法模型集成为强算法模型，可以用于解决分类问题也可以用于解决回归任务[4]。随机森林的出现主要是为了解单一决策树可能出现的很大误差和过拟合的问题。算法的核心思想是将多个不同的决策树进行组合，利用组合降低单一决策树有可能带来的片面性和判断不准确性。通过随机森林算法可以从小区信息表里提取出与流量相关的关键特征，为后面流量预测提供训练特征。侧重于对海量特征提取并排序，特征选择算法如下。

在数据集D中，根据某个特征A的信息增益进行特征选择，信息增益选择算法：

$$G(D,A)=H(D)-H(D/A) \tag{1}$$

数据集D的经验熵为：

$$H(D)=-\sum_{K=1}^{K}\frac{|C_K|}{|D|}\log\frac{|C_K|}{|D|} \tag{2}$$

其中C为输出的无线故障分类。

设有K个类C_K，K=1,2…K，$|C_K|$为属于类C_K的样本个数，有：$\sum|C_K|=|D|$

条件熵：

$$\begin{aligned}H\left(\frac{D}{A}\right)&=-\sum_{i,k}p(D_K,A_i)\log_p(D_K/A_i) \\ &=-\sum_{i=1}^{n}\sum_{k=1}^{k}p\left(\frac{D_K}{A_i}\right)\log_p(D_K/A_i) \\ &=-\sum_{i=1}^{n}p(Ai)\sum_{k=1}^{k}p\left(\frac{D_K}{A_i}\right)\log_p(D_K/A_i)\end{aligned} \tag{3}$$

$$= -\sum_{i=1}^{n}\frac{D_i}{D}\sum_{K=1}^{K}\frac{|D_{ik}|}{|D_i|}\log\frac{|D_{ik}|}{|D_i|}$$

最终选取出影响流量的关键特征为PRB利用率、RRC连接数、X2切换次数、上行流量、下行流量。

2.3.2 LSTM算法模型

LSTM是一种带记忆功能的神经网络，在原来的循环神经网络基础上增加了控制门对数据进行记忆或舍弃[5]。如图4所示，把随机森林选取的流量特征作为输入向量，通过多层神经网络的处理来预测未来的流量。侧重于预测平稳数据，准确度较高。

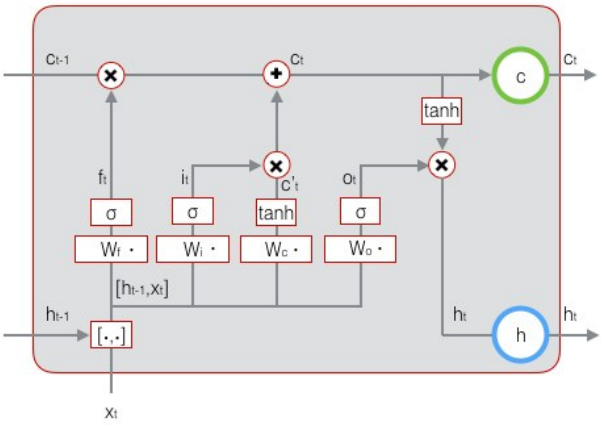

图4　LSTM神经网络模型图

遗忘门：用于基站流量数据序列中需要遗忘的信息。

$$f_t = \sigma\left(W_f\right).\left[h_{t-1},X_t\right] + b_f \tag{4}$$

输入门：用于本次基站流量数据序列的输入信息。

$$i_t = \sigma\left(W_i\right).\left[h_{t-1},X_t\right] + b_i \tag{5}$$

$$\widetilde{C}_t = \tanh\left(W_c\right).\left[h_{t-1},X_t\right] + b_c \tag{6}$$

状态更新：将旧的单元状态C_{t-1}更新为新的单元状态C_t。

$$C_t = f_t{*}C_{t-1} + i_t{*}\widetilde{C}_t \tag{7}$$

输出门：最后需要决定输出内容，也就是要预测的小时级、天、周级的流量数据。长短期记忆输出门的输入为当前时间步输入与上一时间步隐藏状态，输出门由激活函数为Sigmoid函数的全连接层计算得到，然后再通过输出门来控制从记忆细胞到隐藏状态的信息流动，具体实现方式是将单元状态置于tanh（将值推到介于-1和1之间）并将其与Sigmoid门的输出相乘[6]。

$$\sigma_t = \sigma\left(W_o.\left[h_{t-1},X_t\right] + b_o\right) \tag{8}$$

$$h_t = O_t{*}\tanh\left(C_t\right) \tag{9}$$

LSTM主要涉及到的参数：rnn_unit（神经元）=50；input_size（输入大小）=9；output_size（输出大小）=1；lr（学习率）=0.001；batch_size（每次训练大小）=10；time_step（时间延迟）=10；1000个训练集；200个测试集；训练600步。

如图5所示，对基站小时粒度数据进行预测，可以明显地看到该小区的数据在50时刻处发生了突然增大然后迅速回落至较低值，这种现象较为普遍，一般称之为“突发事件”引起的流量变化，这种突变情况难以预测，是本课题的一大难点所在。图5所示的只是预测下一个时刻的情形，其准确度较为可观。但是在突变流量处预测精度欠佳，只能把握其趋势，不能准确地预测到其数值大小。后续会对天、周、月的峰值数据做训练，看是否可以把峰值流量预测精度提升。

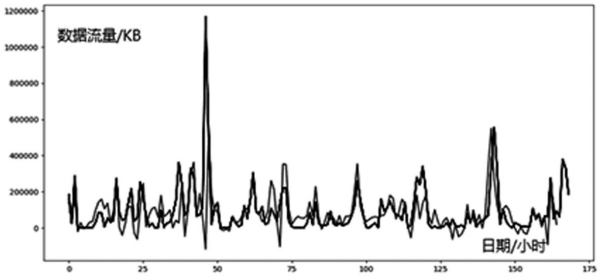

图5　LSTM模型基站流量小时级预测效果图

2.3.3 ARIMA算法模型

ARIMA自回归移动平均模型是指将非平稳时间序列转化为平稳时间序列，然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型[7]。对周期性数据预测较准确，这里设置p=2，q=3。

$$\text{自回归模型：(AR)}y_t = \mu + \sum_{i=1}^{p}\gamma_i y_{t-i} + \varepsilon_t \tag{10}$$

y_t是当前值，u是常数项，p是阶数，γ_i是自相关系数，ε_t是误差，y_{t-i}为前几天的值。

移动平均模型：

$$(\text{MA})y_t = \mu + \sum_{i=1}^{p}\theta_i y\varepsilon_{t-i} + \varepsilon_t \tag{11}$$

$$\text{自相关系数：}ACF_{(K)} = \rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \tag{12}$$

取200小时数据做预测验证，在未加偏置之前的数据展示如图6所示。粗线为原始真实数据，细线部分为预测数据。总体来看，预测精度欠佳，但是其对趋势预测十分精准。

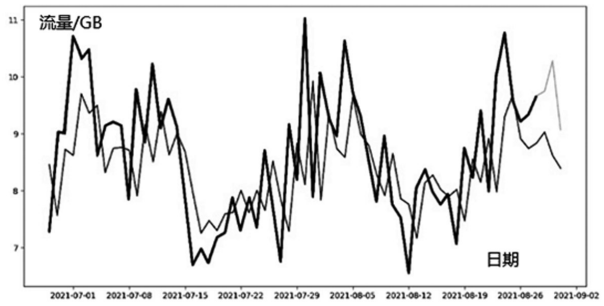

图6　ARIMA模型基站流量预测效果图

3 结果与分析

采用Stacking集成算法对预测模型进行优化，如图7所示。

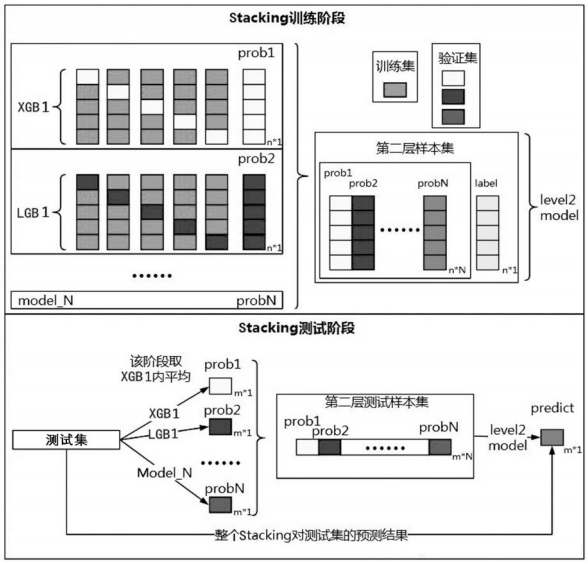

图7　Stacking集成算法对预测模型进行优化过程图

首先是Stacking训练阶段，用多个基础模型分别进行k折交叉验证，图中第一层共有N个基础模型，第1个基础模型为XGBoost，记作XGB1，作为基础模型Model1，第2个基础模型为lightGBM，记作LGB1，作为基础模型Model2，将Model1到ModelN分别进行5折交叉验证，具体来说就是将训练数据平均分为5份，其中4份作为training data，另外1份作为验证数据。每一次的交叉验证包含两个过程：一是基于training data中的4份数据训练模型；二是基于训练生成的模型对验证数据进行预测[8]，然后得到每个模型k折交叉验证预测结果，将结果拼接得到模型对整个训练数据的预测结果，依此类推，得到各个模型的预测数据prob1到probN。将第一层得到的prob1到probN做为第二层模型的训练数据样本集，lable标签数据不变，让下一层的模型基于此进一步训练，训练第二层模型，完成整个Stacking训练阶段。

然后是Stacking测试阶段，将测试数据集分别输入到训练阶段第一层的各个基础模型中，这里需要注意，因为第一层的基础模型采用了k折交叉验证，所以每一个基础模型的预测值为k折交叉验证模型的平均值。同训练阶段一样，将第一层基础模型的输出值prob1到probN进行拼接做为第二层模型的输入样本集，这样经过第二层模型预测输出，就得到了测试样本集的预测值，完成整个Stacking测试阶段。

衡量预测准确性采用SMAPE指标，指标低于1说明预测效果较好，大于1说明预测效果较差：

$$\mathrm{SMAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\frac{\left|\hat{y}_i-y_i\right|}{(\left|\hat{y}_i\right|+\left|y_i\right|)/2} \tag{13}$$

SMAPE平均值项：上行流量为0.559559344，下行流量为0.513604892。图8是上下行基站的流量与真实流量的对标效果图。

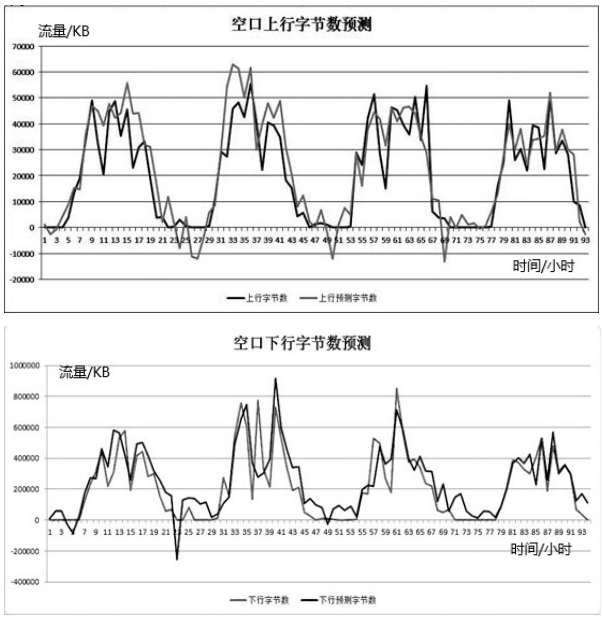

图8　上下行基站流量天粒度最终预测效果图

对流量突变值不显著的基站进行预测，从结果上可以看出与预测结果比较吻合，说明LSTM模型与ARIMA模型对于基站流量的预测是有很大参考意义的。

4 结束语

周期性的数据利用ARIMA和LSTM神经网络模型预测的效果是比较好的，但是对于异常突发流量很难预测，后续可以对重大节假日等特殊日期进行标注，做注意力模型，发现异常流量数据，同时分析用户更新套餐等行为对流量产生的影响来逐步完善基站流量预测模型的准确性。

（下转第29页）

4.4 多场景安全事件响应与处置

基于多年的重保支撑、应急演练、日常运维等多种场景下的实战经验，平台提供面向多场景安全事件处置技术手段来提高安全事件的处置效率，具体如表4所示。

表4　安全事件处置技术手段表

技术手段	主要功能
微隔离	自动采集业务网络中的流量信息，分析并固化业务之间的调用关系，识别出资产与资产间异常流量情况，支持通过预警和阻断等多种方式实现异常流量的快速隔离
一键封停	通过主机代理，快速地对主机上的异常进程或服务下达指令，关闭或停止有安全风险的资产或服务
一键派单	通过接口将安全事件派发给其他周边安全事件处置系统（如：安全运营管理平台SOC、安全事件编排与响应平台SOAR等），完成安全事件的闭环处置

5 实战应用案例

由于资产管理涉及的部门多、分布广、变化快，部分资产老旧和历史原因，很难通过管理手段做到100%准确，长久以来对于变化的资产管理及“三无七边”管理工作一直是运营商在安全运营过程中的核心痛点。

网络空间资产测绘与攻击面管理平台在运营商的大网环境中充分进行了能力验证。在某省级运营商项目实战中，采用主被动融合探测技术，构建了全面、准确、动态的资产测绘能力，资产覆盖度超过98%；结合运营商的海量数据，平台积累了10万+的资产指纹数据和2万+的企业指纹数据，通过运用智能化的指纹识别引擎，资产识别准确率94.6%；平台对授权资产围绕漏洞、安全基线、异常账号等脆弱性，开展风险自动化持续监测，累计了1200多项行业基线核查数据和1万+安全加固解决方案库；在实际的安全运营支撑中，并非所有漏洞都需要解决，有些漏洞无论如何都会持续存在[6]，通过评估漏洞确定漏洞处置的优先级，辅助安全运营工作有序、高效地开展，抓住关键，用20%的时间解决80%的网络安全风险；另外，针对不同的应用场景，平台还提供了多种快速安全事件处置手段，提升安全运营效率，帮助该省级运营商实现了安全运营从“被动防御”向“主动防御”进阶。

6 结束语

网络空间资产测绘与攻击面管理为企业内IT基础设施中的资产与安全风险提供持续的从发现、梳理、监控到治理一站式的管理能力，要求比攻击者更快获悉安全风险，让安全主动防御体系具有敏锐的感知力、精准的预判力、对网络攻击的及时阻断力和对攻击处置后的可追溯力[7]。虽然本研究在网络空间资产测绘与攻击面管理上积累了一定的技术和业务能力，但针对IPv6的海量地址的资产探测、数据资产的测绘等一些新领域上仍存在不小的挑战，这也将会是网络空间资产测绘与攻击面管理下阶段持续研究的方向。

【参考文献】

[1]廉新科，闫卿.基于攻击面的安全评估体系研究[J].通信技术，2020(10):2567-2572.

[2]沈传宝.从漏洞管理到攻击面管理[J].中国信息安全,2022(6):60-62.

[3]黄康宇,杨林,徐伟光,张涛,李华波.软件系统攻击面研究综述[J].小型微型计算机系统,2018(8):1765-1773.

[4]新华社.习近平：在网络安全和信息化工作座谈会上的讲话 [EB/OL]. http://www. gov. cn/xinwen/2016-04/25/content_5067705.htm，2016-04-25.

[5]张人杰,曾振,肖玮.网络安全实战攻防演练部署研究[J].湖南邮电职业技术学院学报,2019(3):23-25.

[6]段铁兴.企业攻击面管理的7个实践[J].计算机与网络，2021(13):50-51.

[7]赵姗.网络安全主动防御体系浅析[J].网络安全技术与应用,2022(4):4-5.

（上接第8页）

【参考文献】

[1]张敏，高科，杨凌云.5G网络分流比提升研究[J].湖南邮电职业技术学院学报，2022(1)：1-4.

[2]舒培炼，刘正兴，史大军，张敏.VoLTE丢包分析与特性参数优化研究[J].湖南邮电职业技术学院学报，2020(2)：8-12.

[3]胡漾.基于模糊理论的5G异构网络切换算法研究[J].湖南邮电职业技术学院学报，2021(2)：13-17.

[4]上海大唐移动通信设备有限公司.基于随机森林分析VoLTE网络故障原因的方法及装置：CN201810444550.0[P].2018-05-10.

[5]徐运武.5G CQI指标优化方案的应用研究[J].湖南邮电职业技术学院学报，2022(3)：8-12.

[6]张文刚.基于深度学习的交通预测技术及其在通信中的应用研究[D].成都：西南交通大学,2018.

[7]中国移动通信集团浙江有限公司.VoLTE网络故障检测方法及系统:201810981353.2[P].2018-08-25.

[8]康丁文.5G通信系统中高效LDPC译码技术研究[D].西安：西安电子科技大学,2019.