

基于机器学习方法 的设备监控指标预测

Device Monitor Prediction Based on Machine Learning Methods

石旭荣¹, 陈辰², 刘权¹, 高有利¹, 刘贤松¹ (1. 中国联通网络AI中心, 上海 200050; 2. 上海诺基亚贝尔股份有限公司, 上海 201206)

Shi Xurong¹, Chen Chen², Liu Quan¹, Gao Youli¹, Liu Xiansong¹ (1. China Unicom Network AI Center, Shanghai 200050, China; 2. Nokia Shanghai Bell Co., Ltd., Shanghai 201206, China)

摘要:

主要介绍了基于机器学习方法的设备监控指标的模型设计与实现。基于时间序列算法,从生产实际出发,在不同应用场景下对不同厂商,甚至是同一厂商内不同设备,不同版本,设计和构建了“冷启动”的指标预测模型。使得指标的预测“千人千面”,并同时达成数据闭环和模型更新的效果。在应用上主要用于监控不同维度的网络业务恶化的问题。

关键词:

设备;时间序列;预测;数据闭环;AlaaS

doi:10.12045/j.issn.1007-3043.2020.10.004

文章编号:1007-3043(2020)10-0015-06

中图分类号:TN915

文献标识码:A

开放科学(资源服务)标识码(OSID):



Abstract:

It mainly introduces the model design and implementation of device monitoring value predict based on machine learning methods. Based on the time series algorithm, from the actual production, it designs and builds a "cold start" indicator prediction model for different manufacturers, even the same manufacturers in different devices, different versions under different application scenarios, which makes the indicator's prediction diverse, and achieves the effect of data closed loop and model update at the same time. It is mainly used to monitor the deterioration of network performance in different dimensions.

Keywords:

Device; Time series; Predict; Data closed loop; AlaaS

引用格式:石旭荣, 陈辰, 刘权, 等. 基于机器学习方法的设备监控指标预测[J]. 邮电设计技术, 2020(10):15-20.

0 前言

在通信网络中,不同设备商的设备其自身在运行工作时,通过对设备进行如性能和可靠性等的监控会形成一系列的指标数据,设备侧的异常是引起业务质量恶化和业务故障的源头之一,借助人工智能,基于这些监控数据进行指标的预测,从而在设备端提前进行预警并采取行动预案,以降低业务的影响度。

本文将探讨从实际情况出发基于机器学习进行设备监控指标异常预测的工作,旨在快速发现网络业

务质量恶化的问题,做到先于用户发现问题。

1 问题分析

不同于其他 KPI 数据,从过往设备监控数据的获取过程中发现,引起业务质量恶化的根因有很多,设备端的异常也仅仅是众多原因中的一条,很难确定业务质量恶化和设备监控信息的具体数据对应关系。同时不同场景下,设备的监控指标也会有不同的表现形式,此外不同小区,对应着不同厂商的设备,这就使得数据本身很复杂。

这也意味着,如果从有监督学习来设计模型的方式的阻力会很大,并且有标签的数据量非常少,不足以

收稿日期:2020-09-01

支持有监督模型的训练学习。

2 解决方案

2.1 模型方案

在进行方案设计的过程中,必须考虑到上述的这些实际情况。从数据的特点上来看,设备运行时的监控指标是典型的时间序列数据。这样可以通过时间序列算法来进行基础模型的选型。对时间序列的数据来说,可以设置时间窗口和窗口长度,确定一个 T 时间点,使用 $T+0$ 之前的数据进行训练,使用 $T+1$ 及之后的数据进行测试,同时可以调整窗口的长度 N ,预测未来 N 的时间长度的指标数据,这样模型训练中训练集和测试集数据的问题可以得到解决。

同时再结合场景,根据不同场景进行时间序列模型的构建,会使得模型更具有针对性。进一步的,可以在不同场景中继续区分不同的设备厂商甚至是同一个厂商不同设备的类型和不同的设备版本,这样会使模型更加聚焦。如此,“冷启动”这个模型时可以获得精度上的保证,并对后续的数据沉淀和模型优化更新积累质量比较高的“打标”数据。

图1示出的是模型方案。

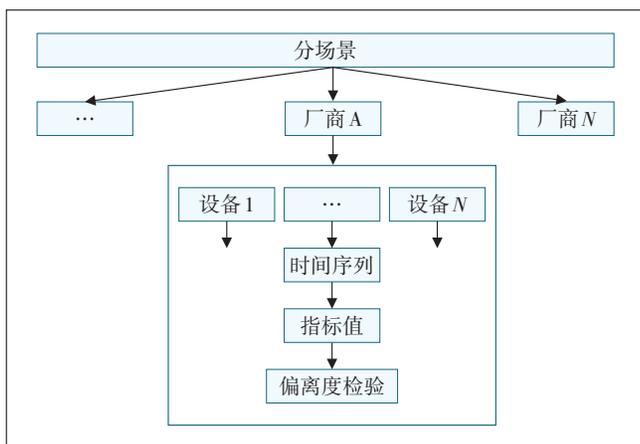


图1 模型方案

2.2 模型构建

在设计好模型的方案后,根据模型方案,分步进行模型的构建与实现,其中机器学习算法的选型是关键。

在同一区域的数据中,不同场景的分离可以通过结合B域数据,O域数据来进行实现,并找到合适的场景类别数。这方面的工作可以通过聚类算法实现。在此次实现中,通过对不同聚类算法的比较并进行选择,最后使用了K-Means聚类算法进行场景的分类,

并通过聚类算法的肘分析法来确定场景分类的具体数量。

也可以结合上述数据对厂商及设备直接进行区分。在实际的情况中,有的厂商,其设备类型或者设备版本对应的数据量很小,则不对该厂商做进一步区分,而是直接建模。这样,不同厂商对应不同的厂商模型,同时不同厂商下也会存在不同的设备模型。

而这些模型也可以根据数据的特点分别使用不同的时间序列算法来进行实现,可以选择如下的时间序列模型。

回归模型:把监控的指标值和时间做对应关系,通过回归模型来做值的预测。回归模型主要包括线性回归、逻辑回归、多项式回归、逐步回归、岭回归、套索回归、ElasticNet回归等。在此次应用中,绝大部分网络性能指标会随着用户行为周期性而发生变化,在多变量预测的情况下虽然回归模型有较好的解释性,但是无法兼顾到指标周期性变化的特征,所以具有周期性的业务质量指标不建议用回归模型进行预测。

ARMA/ARIMA等模型:ARMA自回归滑动平均模型是研究时间序列的重要方法,由自回归模型(简称AR模型)与移动平均模型(简称MA模型)为基础“混合”构成。ARIMA模型名为差分整合移动平均自回归模型,又称整合移动平均自回归模型(移动也可称作滑动),也是时间序列预测分析方法之一。如果设备的监控时间序列数据是稳态的,时间序列也不会随着时间的推移而出现非稳态的变化,那么在此场景下可使用这类模型进行未来值的预测。

时间序列分析模型:是定量预测方法之一。它包括一般统计分析(如自相关分析、谱分析等),统计模型的建立与推断,以及关于时间序列的最优预测、控制与滤波等内容。经典的统计分析都假定数据序列具有独立性,而时间序列分析则侧重研究数据序列的互相依赖关系。后者实际上是对离散指标的随机过程的统计分析,所以又可看作是随机过程统计的一个组成部分。故而,时间序列分析模型也经常用于异常检测。

基本思想:根据系统的有限长度的运行记录(观察数据),建立能够比较精确地反映序列中所包含的动态依存关系的数学模型,并借以对系统的未来进行预报。

基本原理:一是承认事物发展的延续性。应用过去数据,就能推测事物的发展趋势。二是考虑到事物发展的随机性。任何事物发展都可能受偶然因素影响,为此要利用统计分析中加权平均法对历史数据进

行处理。

模型结构: 时间序列 $F(x)$ 可以表示为 3 个因素的函数, 即:

$$F(x) = T(x) + S(x) + R(x)$$

图 2 示出的是原始指标趋势图。

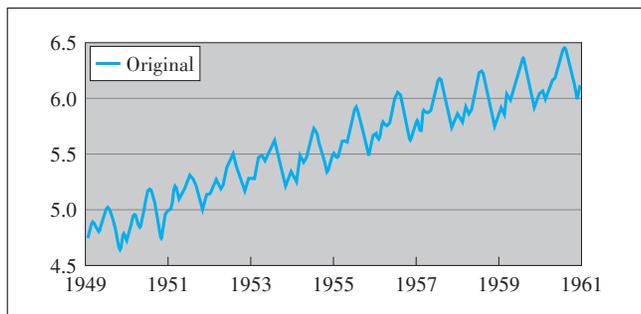


图 2 原始指标趋势图

长期趋势: $T(x)$ 为趋势, 长时间的、缓慢的变化 (见图 3)。

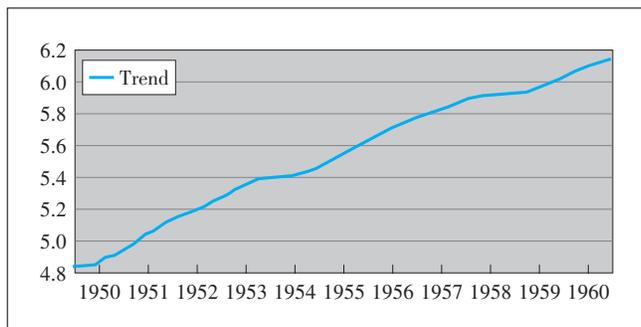


图 3 分解后 $T(x)$ 趋势图

周期性趋势: $S(x)$ 为季节、周期, 周期性的变化 (见图 4)。

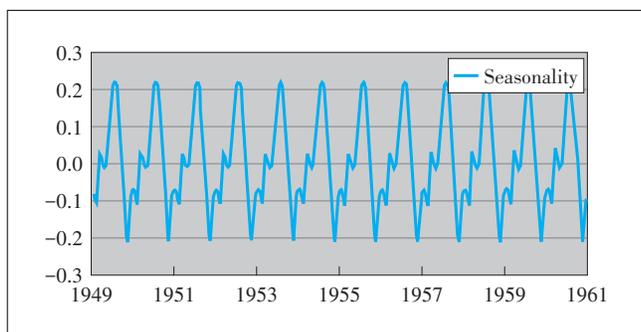


图 4 分解后 $S(x)$ 趋势图

随机性变化: $R(x)$ 为随机性变化, 即各种原因导致的系统变化 (见图 5)。

时间序列分解的方法有很多, 较常用的模型有加法模型和乘法模型。

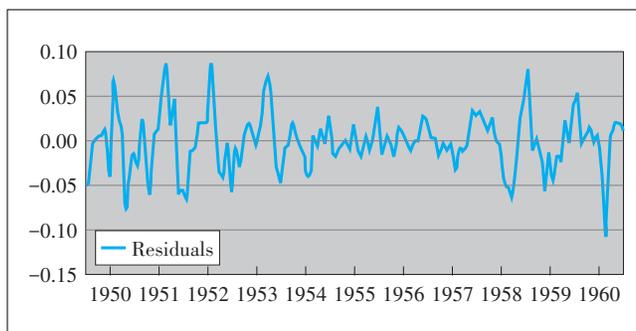


图 5 分解后 $R(x)$ 趋势图

加法模型为: $F(x) = T(x) + S(x) + R(x)$; 乘法模型为: $F(x) = T(x) \cdot S(x) \cdot R(x)$ 。

加法模型更多倾向于组成模型的各个因子是独立的, 乘法模型更多倾向于各个因子之间有影响, 尤其是周期性变化比较明显的应用。由于告警通常是多种因素导致的, 因素与因素之间基本独立, 所以, 理论上本算法适合使用加法模型。但实际应用中发现, 某些指标 (如流量) 数值变化特别大, 单独使用加法模型易受数值范围的影响, 于是对数据取对数, 并加上合适的偏移量, 最终将加法模型转换为有偏置的乘法模型。

通过时序分解, 可对单一指标过去 n 周的数据分解为 3 个函数, 其中趋势和周期是由系统决定的 (稳定), 随机项是应该特别关注的, 正常情况下, 随机项的范围应该处于一定合理范围内, 一旦超出合理范围则表示异常。这里假设异常服从正态分布 (当因素较多、互相独立时这种假设比较容易验证), 根据正态分布密度函数推出。图 6 示出的是正态分布图。

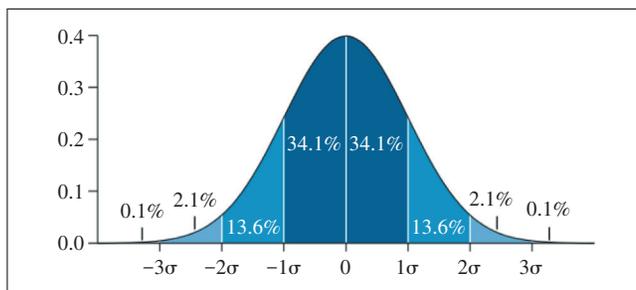


图 6 正态分布图

处于 1σ 之内占总体的 68.3% 而处于 3σ 之内则占到整体的 99.7%, 实际上关注的是 3σ 之外的分布 (实际应用中通常取 $3.5\sigma \sim 4.5\sigma$), 这部分理论上不应该发生, 一旦发生即表示需要特别关注。所以

$$P(x) = T(x) + S(x) \pm 3\sigma$$

输出为预测门限的上限和下限, 以 N 个周期为单位, 提前预测 N 天的告警门限 (其中 N 可以自定义, 当

前 $N=7$)。

在对设备的监控应用中,时间序列分析模型适用于大部分业务质量的恶化监控,其中加法模型适用于指标波动较大的指标,如业务流量等;乘法模型适用于指标波动较小的指标,如成功率、异常占比等。

Facebook-prophet:该算法基于时间序列分解和机器学习拟合实现,可以设置突变点,以及对季节性和趋势性进行预测,亦可以考虑节假日的影响,对有些场景来说,比如在已知确定的节假日时间,重大事件活动时间的情况下,使用该模型会更有优势。

模型结构:Prophet的本质是一个可加模型,基本形式如下:

$$y(t)=g(t)+s(t)+h(t)+\epsilon t$$

其中 $g(t)$ 是趋势项, $s(t)$ 是周期项, $h(t)$ 是节假日项, ϵt 是误差项并且服从正态分布。Prophet算法使用加性模型的好处是可以刻画时间序列的不同变化部分:整体走势、周期性变化、节假日影响和异常点。

整体走势:Prophet里使用了2种趋势模型:饱和增长模型和分段线性模型。这2种模型都包含了不同程度的假设和一些调节光滑度的参数,并通过选择变化点来预测趋势变化。

周期性变化:Prophet用傅里叶级数来建立周期模型。

$$s(t) = \sum_{n=1}^N \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right)$$

对 N 的调节起到了低通滤波的作用。

节假日影响:节假日需要用户事先指定,每一个节假日都包含其前后的若干天。比如中国的双十一、春节以及美国的黑五等。

异常点:拟合那些无法用模型描述的异常点。

所以,如果设备的指标数据不具有明显的周期性,也没有很强的趋势性,那么该算法并不是首选。但大部分业务质量指标都具有明显的周期性,而且也同样会受到节假日的影响。故而,对于网络质量恶化的监控来说,这个算法只需要做少许的修改就能使用网络质量恶化的监控,而且也是落地性较强,效果较好的算法。

LSTM:LSTM通常被称为“长短期记忆网络”,是一种特殊的RNN,能够学习长期依赖关系,属于一种基于时间序列的深度机器学习算法。LSTM是一种含有LSTM区块或其他的一种类神经网络,与RNN神经元一样,LSTM神经元在其管道中可以保持记忆,以允许

解决顺序和时间问题,而不会出现影响其性能的消失梯度问题,也是在训练数据量较大情况下会选择的算法。在此次课题中,出现某个厂商可以获取1年以上数据量的情况,使用了该算法,具有较好的性能,但当数据量较少,在半年以下时,该算法的表现反而不如其他模型,极端情况下甚至也不如回归模型。

图7为该模型在某个厂商数据上的模型表现,设置预测的窗口长度在3时,亦显示了较高的预测精度。

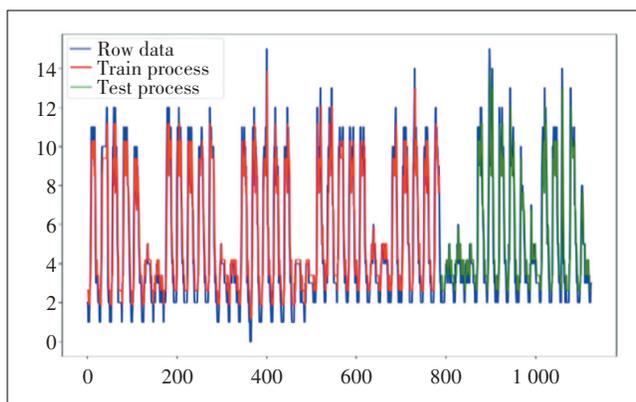


图7 厂商设备监控指标LSTM模型性能

以上的模型需要结合实际情况以及业务的特征进行选择,不同的业务场景、不同的数据特征、不同的数据条件都会影响最终的算法选择。在未来,可以考虑结合知识图谱的方式,学习算法与数据特征之间的关系,最终实现算法的自动选择。

2.3 模型的更新

在得到“冷启动”的模型后,通过AIaaS进行模型的部署上线,并和生产系统进行对接生产数据。同时时间序列模型在 T 时间点对 $T+1$ 时间数据的预测,在 $T+1$ 时刻到来时即可能得到验证,故而能快速地形成模型和数据的闭环。

如此可以通过AIaaS平台定时任务或周期性地发起对模型的重新训练和优化,进行模型的更新。达到数据闭环和模型闭环的效果。如此在训练阶段即使准确度较低的“冷启动”模型也能够在灰度发布后,通过在线试运行一段时间以适配提高模型精度。

图8示出的是线上数据的闭环和模型的更新。

3 模型案例分析

3.1 小区感知监控应用案例

受用户行为因素影响,小区感知指标的波动性很大,并且具有周期性的特征。所以,对于小区感知指标

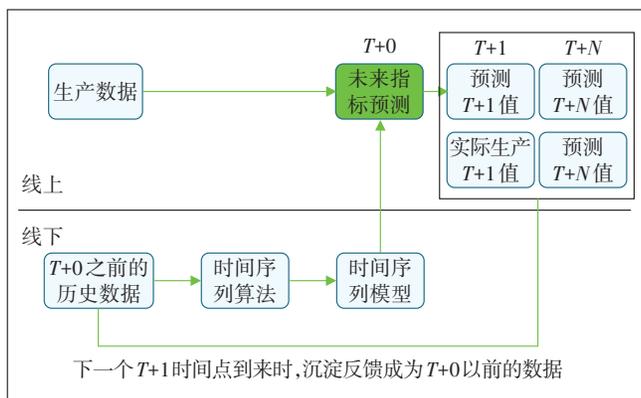


图8 线上数据的闭环和模型的更新

恶化问题的监控,在上述算法中Prophet是最为合适的算法。在本案例中将采用Prophet算法对重要区域的小区进行准实时的感知恶化的监控,以缩短问题发现时间,保障用户感知。

例:小区19*****9, TCP时延连续恶化。小区指标趋势如图9所示。

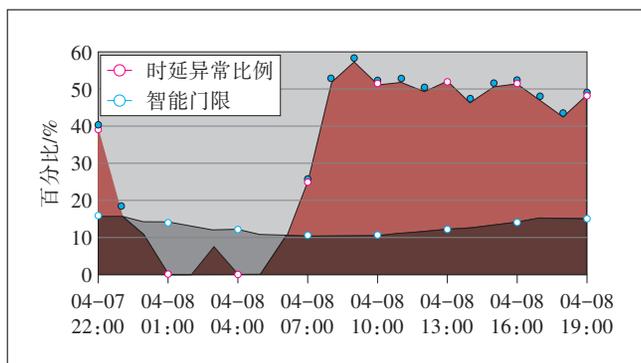


图9 小区指标

图9中灰色的曲线代表预测的指标恶化阈值,红色代表实际的指标值。其中指标恶化阈值是通过Prophet算法计算得出的,即图9中灰色的阴影部分是指标正常的波动区间,一旦指标超出这个区间,那么就是一次用户感知的恶化。在图9中看到15个小区感知恶化的异常点,并且这15个感知恶化的异常时间点影响用户数均在200人以上,其中影响最大的人数为719人。通过快速定位的结果可以看出,这15个小区感知恶化问题都是因为负荷问题导致,无线上行资源受限,最终使得小区下客户的时延恶化,影响客户感知,快速置信度均达到了95%以上,其中近一半的快速定位置信度达到了100%。快速定位亦是机器学习算法得出的,在本案例中采用的是关联算法。通过关联算法从历史数据中得出用户感知恶化与性能指标恶化的

对应关系,从而完成感知指标恶化的快速定位,完成最终的感知恶化监控落地。

经验证,通过Prophet算法识别的感知恶化问题准确率能达到85%。图10为实际验证的汇总结果。

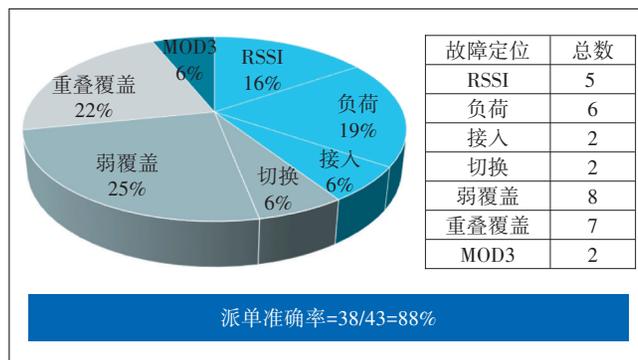


图10 准确率统计图

3.2 全省感知监控应用案例

核心网网元感知指标亦会受到用户行为因素影响,随着活跃用户数的周期性变化,指标也会存在一定的周期性波动,但波动范围较小。在本案例中,采用时间序列分析的算法实现核心网网元感知监控。

图11所示的是某省核心网TCP1/2步握手时延趋势。图11中,灰色的部分为通过时间序列分析算法得出的动态门限,红色的部分为实际的指标趋势。明显可以看出,红色圈出的部分实际指标值已经超过了动态门限的阈值,即为异常的时间点。然后,通过自动的定界,结果显示为NAT以上原因(即多个NAT异常或NAT以上网元的异常)。在本案例中,定界采用的是多维替代分析法。最后通过实际的人工验证证实,网络中有2个NAT分别有2块板子老化导致用户TCP时延异常。

在本案例中,通过核心网的准实时监控,在用户大量投诉前主动发现了问题,并通过及时地处理恢复了正常,实际用户的投诉量与同类问题相比降低至少60%以上,同样证实了采用时间序列的方式进行业务质量监控,可以先于用户发现问题。

目前采用时间序列分析算法的全省级感知监控告警准确性达90%以上。

4 总结

此次基于机器学习进行厂商设备监控指标预测的应用。在进行方案设计的过程中,在分析了监控数据的特点,并从实际的工作角度出发考虑到了不同应用

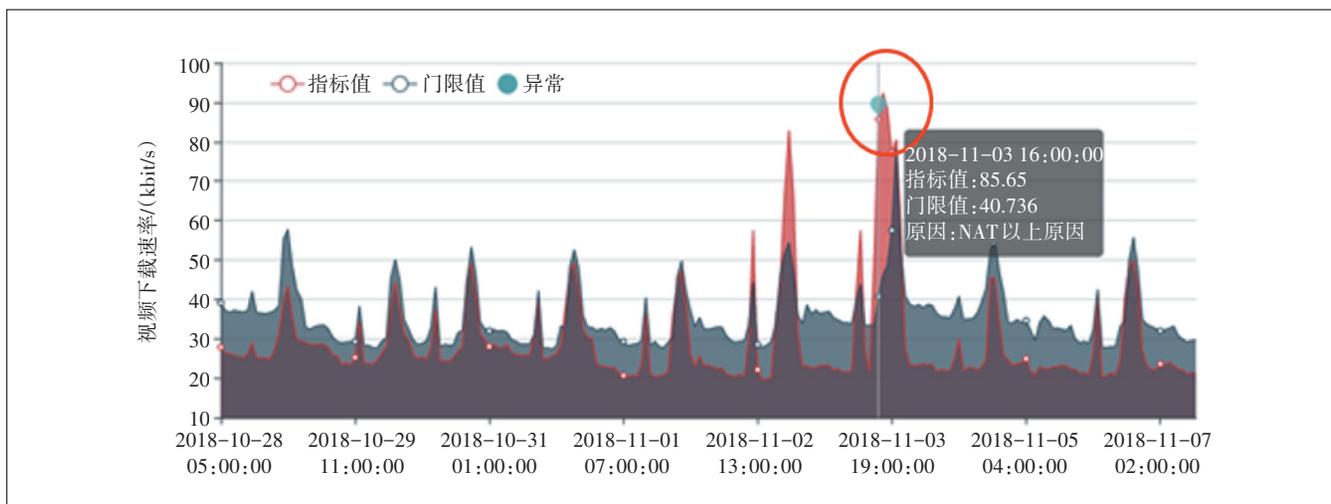


图 11 TCP核心网侧动态门限

场景下不同厂商,甚至是同一厂商内不同设备、不同版本下的情况,设计和构建了“冷启动”的指标预测模型。不同厂商下,对应不同模型,同一厂商下对应不同的设备模型,形成了“千人千面”的特点,上线后,可以通过数据的沉淀继而进行模型的在线优化更新,不断提高模型的在线服务时间。

值得注意的是“基于机器学习进行厂商设备监控指标预测的应用”,只能帮助发现异常,在实际的应用中还需要进行进一步的问题定界定位,才能够最终实现应用落地,并且获得较好的效果。

参考文献:

- [1] pmdarima: Arima estimator for python [EB/OL]. [2020-08-30]. <http://alkaline-ml.com/pmdarima/>.
- [2] JIMMY BA, RICH CARUANA. Do Deep Nets Really Need to be Deep? [C]//International Conference on Neural Information Processing Systems. MIT Press, 2014.
- [3] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
- [4] BACK A, TSOI A. FIR and IIR Synapses, a New Neural Network Architecture for Time Series Modeling [J]. Neural Comp, 1991, 3(3): 375-385.
- [5] YANG J, LIU C, SHANG Y, et al. Workload Predicting-Based Automatic Scaling in Service Clouds [C]//2013 IEEE 6th International Conference on Cloud Computing (CLOUD). IEEE, 2013.
- [6] YANN L, YOSHUA B, GEOFFREY H. Deep Learning [EB/OL]. [2020-08-30]. <https://www.cs.toronto.edu/~hinton/absps/Nature-DeepReview.pdf>spm=a2c4e.11153940.blog-cont576283.17.3ac27677LdbpjU&file=NatureDeepReview.pdf.
- [7] WANG S, JIANG J. Learning Natural Language Inference with LSTM [J]. 2016(8).
- [8] CHENG Y, WANG C, YU H, et al. GRU-ES: Resource Usage Prediction of Cloud Workloads Using a Novel Hybrid Method [C]//2019 IEEE 21st International Conference on High Performance Computing and Communications, 2019.
- [9] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [J]. Computer ence, 2014(9).
- [10] R. J. HYNDMAN, S. MAKRIDAKIS, S. Wheelwright. Forecasting Methods and Applications [M]. John Wiley and Sons, 2018.
- [11] CHENG J, CHEN X, SHEN M. A Framework for Daily Activity Monitoring and Fall Detection Based on Surface Electromyography and Accelerometer Signals [J]. IEEE Journal of Biomedical and Health Informatics, 2013, 17(1): 38-45.
- [12] WILSON G, COOK D J. A Survey of Unsupervised Deep Domain Adaptation [J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(5): 1-46.
- [13] YAROSLAV GANIN, EVGENIYA USTINOVA, HANA AJAKAN, et al. DomainAdversarial Training of Neural Networks [J]. Journal of Machine Learning Research, 2016, 17(59): 1-35.
- [14] JUDY HOFFMAN, ERIC TZENG, TAESUNG PARK, et al. CyCADA: Cycle-Consistent Adversarial Domain Adaptation [M]. In ICML, Stockholm, Sweden, 2018.

作者简介:

石旭荣,毕业于山东大学,工程师,学士,主要从事通信方向的机器学习、大数据分析工作;陈辰,毕业于上海财经大学,主要从事移动网络运维、移动网络优化以及移动网络大数据和AI研究工作;刘权,毕业于布里斯托大学,工程师,硕士,主要从事机器学习、大数据分析工作;高有利,毕业于东北农业大学,高级工程师,学士,主要从事网络应用项目经理工作;刘贤松,毕业于武汉水利电力大学(武汉大学),中国联通网络AI中心副经理,硕士,主要分管网络服务智能化产品研发和网络服务智能化产品推广管理工作。