基于日志的机器学习方法实现故障快速定界的研究与应用

Research and Application of Log-based Machine Learning Method to Realize Fast Delimitation of Faults

杨朝鹏，林业贵，罗飞鹏（中国联通广东分公司，广东 广州 510627）
Yang Zhaopeng，Lin Yegui，Luo Feipeng（China Unicom Guangdong Branch，Guangzhou 510627，China）

摘 要：
主要介绍了基于机器学习的日志分析应用于NFV场景下的故障快速定界定位系统的设计与实现。该系统主要基于大数据分析、人工智能算法、云计算基础设施构建了一个智能化的维护平台，将IT技术的优势应用于CT运维领域，同时继承了CT运维领域故障经验库和规则的知识积累，结合告警和性能指标信息，最大化地实现故障根因的快速定界定位和业务快速恢复。

关键词：
NFV场景；智能算法；大数据；智能化维护平台
doi：10.12045/j.issn.1007-3043.2018.12.005
中图分类号：TP181
文献标识码：A
文章编号：1007-3043（2018）12-0023-04

Abstract：
It mainly introduces the design and implementation of fault-based fast demarcation and positioning system based on machine learning-based log analysis in NFV scenarios. The system is based on big data analysis, artificial intelligence algorithm, cloud computing infrastructure to build an intelligent maintenance platform, which contains the advantages of IT technology in the field of CT operation and maintenance and inherits the knowledge accumulation of fault experience library and rules in the CT operation and maintenance field, combines with alarm and performance indicator information, that can maximizes the rapid demarcation of fault root causes and rapid business recovery.

Keywords：
NFV scenarios；Artificial intelligence algorithm；Big data；Intelligent maintenance and operation platform
引用格式：杨朝鹏，林业贵，罗飞鹏. 基于日志的机器学习方法实现故障快速定界的研究与应用[J]. 邮电设计技术，2018(12)：23-26.

0 前言

设备日志分析作为隐患排查和故障定位的辅助手段，在NFV场景下不可避免地面临诸多问题，主要有NFV场景下层次和部件众多，各部件日志分散，并且日志产生量大，查找困难。这些问题给NFV设备的故障快速定界定位带来挑战。因此需要建立专有日志分析平台，实现跨层日志统一采集、集中存储管理、快速搜索分析、关键日志监控，并通过基于日志的机器学习方法实现故障快速定界定位及基于多种算法实现隐性故障排查。

1 研究背景

1.1 NFV设备部署规模化

随着业务发展的需要，中国联通已开展NFV设备的规模部署。网络功能虚拟化（NFV）作为一种CT与IT进行融合的技术，初衷是为了解决现有发展的瓶颈，解决目前业务创新驱动力不足的问题。NFV是通过软硬件解耦及功能抽象，改变网络设备以往烟囱化的架构，使功能与专有硬件分离，从而使底层资源可以灵活共享，实现新业务的快速开发迭代。中国联通部署NFV设备不仅实现NB-IoT和VoLTE等网络功能，还为将来部署5G网络奠定基础。

1.2 新技术带来维护方法的变革

虚拟化、云化技术使得虚拟计算资源、虚拟网络资源、虚拟存储资源与物理资源的对应依赖关系扑朔迷离。加大了运维人员对电信NFV架构的理解难度，也使运维工作不易开展，网络维护面临着“风险预防难”“事故恢复难”和“根因定位难”三大难题。

大数据分析和人工智能技术的成熟为解决现有维

收稿日期：2018-11-08

护问题提供了新的解决方向。应用大数据分析和人工智能算法，对海量日志数据进行挖掘，可实现NFV场景下典型故障的快速定位，结合告警和性能指标等信息，进一步提高定位准确率。

2 实现原理

传统的故障处理思路是通过告警排查定位显性故障，通过KPI指标、业务信令比对和分析定位隐性故障。随着网络规模的扩大和NFV的部署，传统定界定位方式已不再具备高效性。在研究过程中发现，分析设备日志比分析告警信息能更敏锐地探测到系统隐患和设备故障。但是海量的设备日志，降低了排查效率。为解决人工排查效率低下的问题，需要应用人工智能技术到日志分析过程中，利用数据挖掘快速实现基于日志的故障诊断。考虑到排查的效率，对能提供故障特征、标记、匹配算法和机制的故障，视为已知故障，导入特征库（专家经验库），用于故障快速匹配。特征库数据可通过人工标注和专家经验导入。对于特征库未能匹配的故障，视为未知故障（状态未知），可通过机器学习等方法分析日志分布规律的变化，完成异常检测和异常根因分析，实现基于日志的未知故障诊断。

图1示出的是日志分析实现原理。

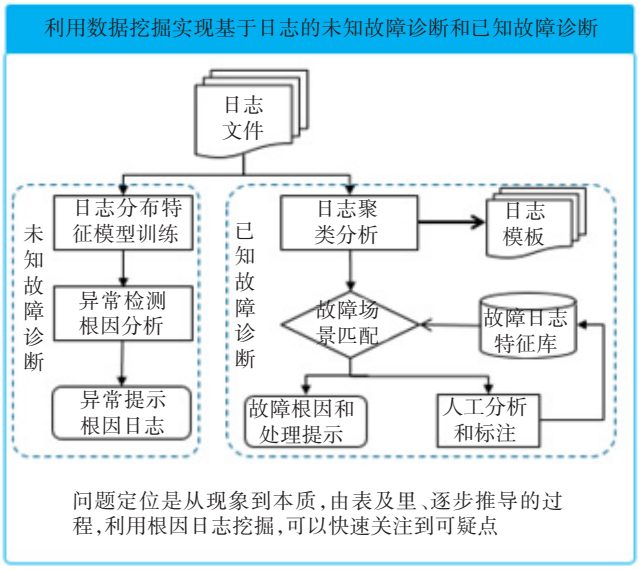

图1 日志分析实现原理

在机器学习方面，参考和借鉴了很多应用案例，最终选择STIDE和LSTM算法来进行日志模型的训练和异常检测。STIDE算法作为常用的序列截取方法，通过滑动窗口进行短序列的截取。日志经过模板化处理后，以时间排序，通过STIDE算法，形成不同的集合。如一组指令序列{open、read、move、move、open、move、error}，用滑动窗口依次截取长度K=3的短序列，经过处理后得到的序列分别是P1{open、read、move→move}、P2{read、move、move→open}、P3{move、move、move→error}。

经过处理后的短序列，使用LSTM模型来对日志序列建模，即使用系统正常运行状态下产生的日志序列来训练LSTM模型，让LSTM模型学习到系统正常状态下产生的日志模板，并实现在线的异常检测。

图2示出的是DeepLog架构。

日志关键字异常检测模型（Log Key Anomaly Detection Model）把关键字序列异常检测问题转化为一个多分类问题，即输入一个固定窗口大小的日志模板序列，输出是下一个日志模板的概率分布。例如将截取的序列进行概率分析，P1的概率为0.7，P2的概率为0.2，P3的概率为0.05，将序列按照概率值的大小排序，

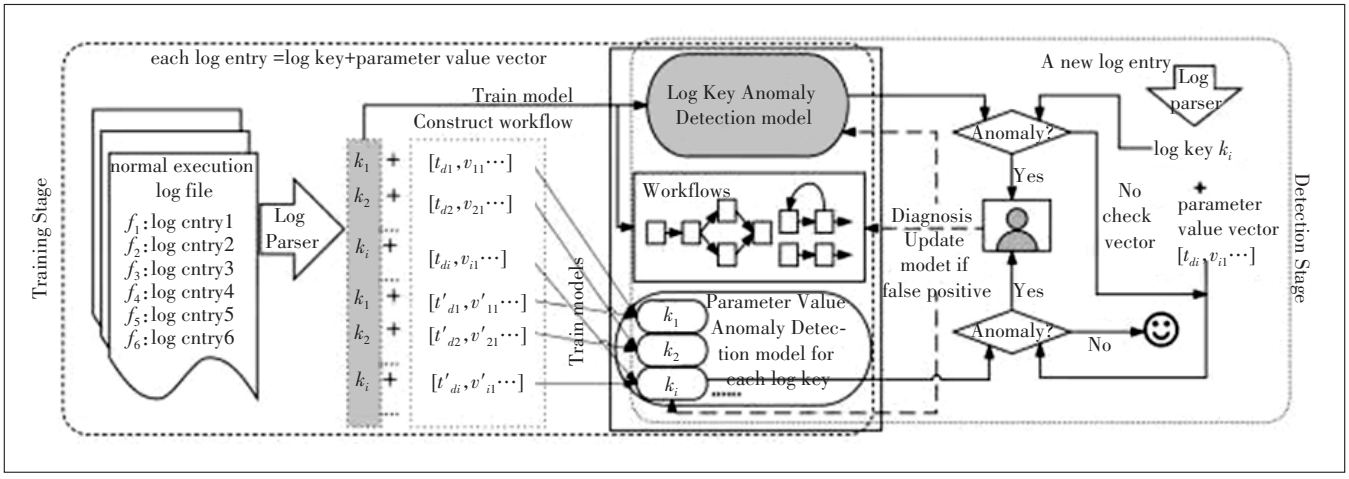

图2 DeepLog架构

取出概率值最大的N个日志模板，将新产生的日志序列X与N个日志模板进行比较，如果结果一致就认为X是正常的日志，如果不一致则是异常日志。即使在日志关键字异常检测中验证通过后，对于日志中的变量信息也有对应的变量异常检测模型（Parameter Value Anomaly Detection Model）进行检测。在变量异常检测训练阶段，把训练数据分为两子集，前一部分归为训练集，后一部分为验证集，使用训练集数据来训练LSTM模型，然后使用训练得到的模型来对模型验证数据集进行预测，得到模型预测值。预测值与验证集中实际向量之间的误差被拟合为一个高斯分布。如果在线异常检测阶段预测得到日志变量向量与真实值之间的均方误差能够位于该高斯分布比较高的置信区间内（比如80%~100%），则认为它是正常的，否则认为它是异常的。

3 解决方案

3.1 系统架构

系统平台部署于PaaS上，基于PaaS的服务生命周期管理、伸缩、高可靠等能力，使其在各层级都可稳定运行和水平扩展。系统平台采集NFV各产品组件、服务、硬件COTS等的日志统一存储进行集中管理和故障定界定位分析。

图3示出的是日志采集与处理架构。

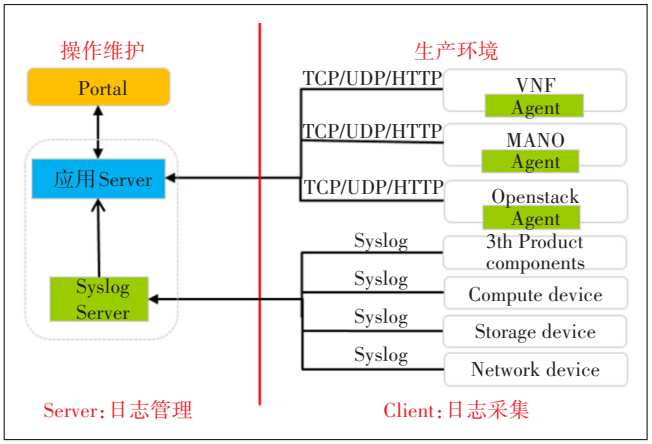

图3　日志采集与处理架构

3.2 日志处理流程

3.2.1 日志预处理

日志预处理主要是将采集的日志基于网络拓扑，根据NFV日志的来源、时间信息和空间维度（包括水平拓扑和垂直拓扑）等要素标注整理入库。

3.2.2 故障处理流程

为实现快速定位，首先对检测日志模型进行专家经验库匹配。匹配不成功再进行机器学习检测，如没有发现异常，最后进行关键字和日志统计异常分析。故障处理流程如图4所示。

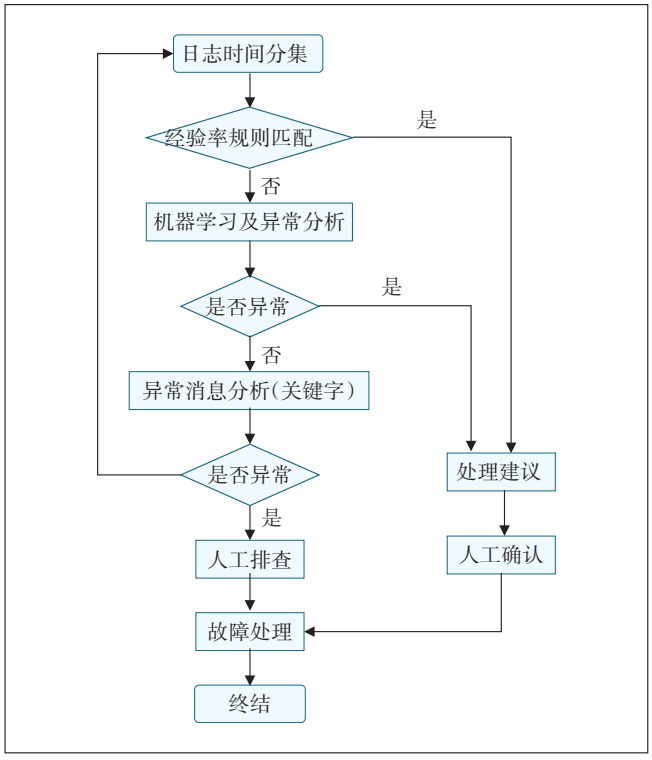

图4　故障处理流程

3.3 关键组件

3.3.1 故障经验库以及规则整理

故障经验库用于显示和管理当前系统中的所有故障诊断规则，用于现网故障分析；通过整理历史故障案例，可以定义大部分常见故障规则，后期也可以通过结果反向更新规则。故障诊断规则包含故障现象、相关的根因描述、处理建议以及识别规则。故障识别时，会根据用户添加的规则及根因进行识别。

3.3.2 机器学习与异常检测

作为系统核心组件，机器学习与异常检测组件基于行业成熟应用的大数据深度学习预测引擎。在系统部署初期需要大量原始数据进行训练，首先进行正常业务日志模型学习，通过文本聚类算法，得到特征字，并形成特征模板。日志通过特征模板训练，再通过STIDE算法截取，构成业务模型。对大量业务模型进行概率运算，构建日志模板异常检测模型库，以概率排序，概率高的判断为正常分支。在运行初期，可能会因

为样本不足，出现正常业务模型的概率值不高的情况，这时需要人工标注。

在异常日志检测模式中，设备日志经算法处理后得到待检模型，通过匹配检测模型得到异常分支结果。当然也存在不能匹配到任何检查模型的情况，这时也会显示为异常分支，后续通过人工核查标注后，进入日志模板异常检测模型库。

在实际故障日志诊断过程中，通过对比排查，可以很快分析出异常分支。根据结果信息，进行人工标注自学习分析结果。

人工诊断后可进行以下几种操作。

a）标记为根因：此异常在后续被命中时，将优先显示；该异常支持标记为非根因。

b）标记为非根因：此异常后续将不在界面显示，不支持再次标记为根因。

c）加入经验库：对于能明确问题的异常日志，可加入经验库。

3.3.3 其他异常分析模块

在使用经验库分析和机器自学习都无法查找异常时，可通过消息关键字或日志统计异常进行排查，错误类关键字包括error、fail、failed、fault、abnormal、exception等。

4 结束语

通过基于日志的机器学习研究和实践应用，总结出适合NFV网络的智能分析应用方法。机器学习规则和专家经验规则也在系统运行过程中迭代更新，将有力支撑网络规模发展。另外在应用设计初期，就考虑后续其他开发迭代需求，重点围绕低门槛配置化开发、开放性、云化3个关键点，构筑智能化网络运营能力。因此所有基于智能平台的能力如告警处理、网络诊断、网络/设备集成等，均能以API、页面、数据模型等形式提供原子组件，从而支持新业务和新技术应用快速开发上线，适应未来运维发展。

参考文献：

[1] 史蒂芬·卢奇，丹尼·科佩克. 人工智能[M]. 北京：人民邮电出版社，2018：36-41.

[2] GOODFELLOW I，BENGIO Y，HEATON J，COURVILLE A. 深度学习[M]. 北京：人民邮电出版社，2017：275-298.

[3] 李素游，寿国础. 网络功能虚拟：（NFV架构开发测试及应用）[M]. 北京：人民邮电出版社，2017：83-104.

[4] 顾炯炯. 云计算架构技术与实践[M]. 北京：清华大学出版社，2016：22-37.

[5] DU M，LI F，ZHENG G，et al. DeepLog：Anomaly Detection and Diagnosis from System Logs through Deep Learning[C]// the 2017 ACM SIGSAC Conference. ACM，2017.

[6] 朱洁，罗华霖. 大数据架构详解：从数据获取到深度学习[M]. 北京：电子工业出版社，2016：53-68.

[7] 彭冬，朱伟，刘俊. 智能运维：从0搭建大规模分布式AIOps[M]. 北京：电子工业出版社，2018：74-89.

[8] 祁瑞华. 基于机器学习算法的分类知识发现及其在文本分析中的应用[M]. 北京：清华大学出版社，2015：51-72.

[9] 刘凡平. 神经网络与深度学习应用实战[M]. 北京：电子工业出版社，2018：82-103.

[10] 盛骤，试式千，潘承毅. 概率论与数理统计[M]. 4版. 北京：高等教育出版社，2018：123-138.

[11] YANN L，YOSHUA B，GEOFFREY H. Deep Learning[EB/OL]. [2018-09-30]. https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdfspm=a2c4e.11153940.blogcont576283.17.3ac27677LdbpjU&file=NatureDeepReview.pdf.

[12] JASON BROWNLEE. Stacked Long Short-Term Memory Networks[EB/OL]. [2018-09-30]. https://machinelearningmastery.com/stacked-long-short-term-memory-networks/.

[13] BOWIE C A，DENDY R O，HOLE M J. Delay time embedding of mass loss avalanches in a fusion plasma-oriented sandpile model[J]. Physics of Plasmas，2016，23(10)：100703.

[14] 陈春宝. 大数据与机器学习：实践方法与行业案例[M]. 北京：机械工业出版社，2017：178-192.

[15] ETSI. Network Functions Virtualisation-White Paper #3[EB/OL]. [2018-09-30]. https://portal.etsi.org/Portals/0/TBpages/NFV/Docs/NFV_White_Paper3.pdf.

[16] 陈俊. 基于大数据机器学习技术的IT运营分析系统建设[J]. 计算机时代，2018.

[17] 章思宇，黄保青，姜开达. 统一身份认证日志集中管理与账号风险检测[J]. 东南大学学报：自然科学版，2017(47)：117.

[18] 曹政. 基于Mahout框架的Hadoop平台作业日志分析平台设计与实现[J]. 软件，2015，36(11)：43-47.

[19] 钟雅，郭渊博. 基于机器学习的日志解析系统设计与实现[J]. 计算机应用，2018.

[20] 王满，谢亚楠. 随机森林算法在运营商告警推送中的应用研究[J]. 工业控制计算机，2017(6).

[21] SETA K，IKEDA M . Model based development of a meta-leraning support system to prompt self-awareness through presenation for meta-learning[M]. IEEE，2011.

作者简介：

杨朝鹏，毕业于南昌大学，工程师，学士，主要从事核心网CS域和IMS域技术支持、网络规划以及优化等工作；林业贵，毕业于北京邮电大学，工程师，硕士，主要从事CS域、IMS域核心网维护和建设工作；罗飞鹏，毕业于中山大学，学士，主要从事移动核心网、NFV网络规划、运营管理、分析优化、网络和信息安全等工作。