编号：2023F0145

哈尔滨工业大学
大学生创新训练计划项目验收书

项目名称： 基于机器学习的系统日志异常检测

项目级别： 校级 （国家级、省级、校级）

执行时间： 2023 年 10 月 至 2024 年 9 月

负 责 人： 郁清方 学 号 ： 2022110900

联系电话： 17372154543 电子邮箱： 2022110900@stu.hit.edu.cn

院 系 及 专 业： 计算学部 数据科学与大数据技术

指导教师： 王宏志 职 称 ： 教授

联系电话： 13069887146 电子邮箱： wangzh@hit.edu.cn

院 系 及 专 业： 计算学部 数据科学与大数据技术

哈尔滨工业大学本科生院

填表日期：2024 年 9 月 23 日

一、课题组成员：（包括项目负责人、按顺序）

姓名	性别	所在院	年级	学号	身份证号	本人签字
郁清方	男	计算学部	大三	2022110900	320584200407070015	
鲍健焘	男	计算学部	大三	2022110950	321284200406270212	
卜春元	男	计算学部	大三	2022112840	340222200401221319	

二、指导教师意见：

该项目按计划完成，成果满足计划并有一定创新性，同意参加验收并建议评奖

签　　名：王宏志

2024 年　　9 月　　23 日

三、学院专家组意见：

组长签名：　　（ 盖　章 ）

年　　月　　日

四、项目成果：

其它成果（软件、模型、图纸或作品等）：

序号	名称	说明
1	**系统日志异常检测软件**	含有一个简单的 UI 界面，可通过上载模型参数，随后上传 hdfs 的日志序列进行异常检测；也可使用处理好的数据训练模型，随后进行异常检测。

五、项目研究结题报告

1 课题研究目的

人类社会逐渐步入大数据时代，各大企业、高校也开始建立数据中心来处理海量技术，提高效益或对数据进行各样研究。由于数据中心大多具有计算集群规模庞大、长期处于高强度负载的特性，计算集群硬件故障、系统故障等问题频频发生，若故障不能及时被发现、定位并被排除，可能会对对数据中心的算力、数据吞吐量造成影响，从而影响用户体验和企业收益。

本项目旨在从系统日常运行产生的系统日志，进行日志解析、特征抽取以及异常检测，做到从海量系统日志中快速寻找到异常日志，并向运维人员及时预警，便于运维人员快速定位故障和异常，并快速排除故障，节省人力物力和精力的消耗。

本项目基于机器学习，借助已有的日志解析器对日志进行模版提取，构建基于 LSTM 的日志主题预测模型，利用相关算法将系统日志模板化、提取特征并随后进行异常检测，从而实现从系统日志中筛选出可能存在异常的日志。从而为运维人员提供潜在的故障提示，提高数据中心运行的稳定性并保持高效率运行。

2 课题背景

2.1 背景：

随着网络和新一代通讯技术的高速发展，在线服务需求激增，数据成为关键资源，而数据中心承 担着数据收集、计算、转化、流通的重要职责，日常运行负载始终处于较高状态，加之数据中心通常由成百上千台服务器组成集群，而且随着软件系统越来越庞大和复杂，一个数据中心的各个软硬件模 块每日产生的系统日志数是海量的，运维难度与成本也显著增高。以 HDFS(Hadoop Distributed File System)分布式文件系统为例，该系统可在 38.7 小时的运行时间内，产生超过 11,175,000 条日志记录，并且日志文件的体积可达 1.47GB 。显然在目前的趋势下，单单依靠人工标记分析海量系统日志已经变得不太可能。

2.2 研究历史和现状：

目前针对日志的异常检测可分为三类：基于规则的异常检测方法、基于无监督的异常检测方法、基于有监督的异常检测方法。

（1）基于规则的异常检测：

基于统计学方法通过统计海量日志中的关键字出现频率，结合专家经验设计正则表达式，挖掘潜在检测规则并根据规则匹配程度进行日志异常检测。

（2）基于无监督的异常检测：

不需要预先标记的训练数据，通过判断待检测日志序列与正常日志序列的差异检测异常。

（3）基于有监督的异常检测：

需要使用带有标签的数据训练模型，通过预训练的模型进行日志异常检测。

随着人们对人工智能领域的深入研究以及深度学习的发展促使一些学者将神经网络与日志检测关联到一起，逐渐出现了一些基于神经网络的异常日志检测方法，在实际应用中也取得了较好的效果。

2.3 意义：

运维过程中，相较于其他对象，日志是内容最为丰富且来源充足的一类数据, 系统的异常或者性能下降一般会在日志中优先体现出来。如今体量巨大且种类繁多的日志数据使得人工分析难以满足要求，在探索解决方案的过程中，机器学习成为了研究者们的新选择。

通过机器学习进行日志异常检测，可以迅速的发现其中的故障并对故障进行定位，这对于传统运维方法改进和补充等有积极意义，对促进系统的职能、稳定、全面、安全运行等也有显著价值。

3 课题研究主要内容

3.1 威胁模型

3.1.1 引入原因

在本项目中，我们提出的方法将会从无异常的日志序列中学习到全面而复杂的进程行为模式。

由于训练所用到的日志数据来自互联网巨头的共享，难免存在服务器被攻击而产生的日志，针对这个问题，我们在这里引入两种威胁模型。

我们假设系统产生的日志记录不会被敌对攻击者所篡改，并且敌对攻击者也不能通过篡改系统代码从而使系统无法正常地产生日志记录。因此，我们假设系统日志服务是安全可信的。系统日志服务能够将异常进程行为或者敌对攻击者的行为记录到日志文件内。

3.1.2 引入的两种威胁模型

(1) 敌对攻击导致的系统行为模式产生异常。此类威胁通常是敌对攻击针对系统，从而导致系统产生异常的日志行为模式、服务不断重启以及某些任务由于异常出现而提前中断。

(2) 敌对攻击在系统日志中所留下的痕迹。此类威胁通常是敌对攻击针对系统环境下的某些子任务进程，从而在系统运行过程中留下自己相应的日志痕迹。

3.2 项目研究内容

本项目将对日志采集、日志解析、特征提取以及异常检测四个模块进行学习和研究，并尝试设计实现基于机器学习的日志异常检测。

3.2.1 日志采集

日志是一种信息系统中广泛存在的数据， 往往散落于系统各处， 特别是对于分布式软件、大型软件等，不同节点、不同组件都在持续生成和积累日志，所以日志采集是日志异常检测任务的第一步。较为常见的日志采集工具包括 Logstash、Flume、Fluentd、Kafka 等，通过持续、实时的数据采集，可以将数据集中入库或者通过“发布-订阅”的模式推送给消费端进行后续处理。

通过相关文献我们得知，互联网巨头如阿里巴巴、亚马逊等，它们会将它们的服务器运行产生的部分日志公开在网上，同时对其中异常的日志做了标注，我们可以寻找一份日志，将它分为训练数据集和验证数据集用于后续训练与验证。

在这里我们选择了 HDFS 的日志作为我们的数据集，数据来自 Github（https://github.com/logpai/loghub）[1][2]。该数据集采集于 200 多台亚马逊 EC2 节点，其中涵盖了 575,061 个 block 节点所产生的 11,175,629 条日志记录，其中有 16,838 个 block 节点被 Hadoop 数据领域的专家标记为异常。

3.2.2 日志解析

日志属于半结构化数据，需要分析其一般构成，并利用解析技术将其中的常量和变量分离开来，为后续的特征提取提供良好的基础。

一条日志数据可以分解为正则消息部分和特征消息部分。正则消息包括时间戳、日志等级、产生日志的类名称等日志基本组成 这部分是可以用正则表达式进行提取和分解的， 属于日志数据的基础信息。特征消息部分则是描述了日志内容的核心部分，由文本、数字及特殊符号等组成。

本模块通过日志文本，对日志进行解析，并将日志记录转化为结构化日志，即将日志拆分为日志模板和参数变量。

由于我们选择了 HDFS 的日志作为我们的数据集，所以我们选择了针对 HDFS 的日志解析效果最好的为 Spell 解析器[3][4]。

类别	系统	日志样例	正则消息	特征消息	
				日志模板	参数变量
操作系统	Linux	[OK] Started File System Check on /dev/disk/by-uuid/6d329…5-88652b684ca0.	[OK]	Started File System Check on	/dev/disk/by-uuid/6d329…5-88652b684ca0
数据库	Oracle	ORA-00603:ORACLE server session terminated by fatal error	ORA-00603	ORACLE server session terminated by fatal error	—
硬件	交换机	Nov 9 2018 09:49:03 HuaWei03 %%01SHELL/4/LOGINFAILED(s)[12]:Failed to login. (Ip=192.168.1.199, UserName=admin, Times=3, AccessType=TELNET, VpnName=)	Nov 9 2018 09:49:03 HuaWei03 %%01SHELL/4/LOGINFAILED(s)[12]	Failed to login. (Ip=*, UserName=*, Times=*, AccessType=*, VpnName=*)	(192.168.1.199, admin, 3, TELNET)
分布式系统	HDFS	081109 203519 147 INFO dfs.DataNode$PacketResponder: Received block blk_-1608999687919862906 of size 91178 from /10.250.14.224	081109 203519 147 INFO dfs.DataNode$ PacketResponder:	Received block * of size * from *	(blk_-1608999687919862906, 91178, 10.250.14.224)
Web服务	Apache	[Fri Aug 18 22:36:26 2000] [error] [client 192.168.1.6] File does not exist: /usr/local/apache/bugletdocs/Img/south-korea.gif	[Fri Aug 18 22:36:26 2000] [error]	[client *] File does not exist:	(192.168.1.6, /usr/local/apache/bugletdocs/Img/south-korea.gif)
消息队列	Rabbit MQ	=INFO REPORT==== 3-Jul-2017::11:45:14 === rabbit on node rabbit@node2 up	=INFO REPORT==== 3-Jul-2017::11:45:14 ===	rabbit on node * up	rabbit@node2
应用程序	Health APP	201812-20:5:32:205|Step_ExtSDM|30002312|calculateCaloriesWithCache totalCalories=18289	201812-20:5:32:205|Step_ExtSDM|30002312|	calculateCaloriesWithCache totalCalories=*	18289

图 3.2.2-1 日志解析的样例

3.2.3 特征抽取

这个模块主要在于构造机器学习模型可以处理的特征数据，从而来学习日志的正常或异常模式。所提取的特征质量决定了后期模型检测效果所能达到的精度。

查阅相关文献和技术资料得知，特征提取往往需要首先将日志拆分成不同的小组，每个小组为异常检测的最小单位，而拆分方式主要分为固定窗口(Fixed Window)、滑动窗口(Sliding Window)

和会话窗口(Session Window)。

在我们的项目之中，我们采用了滑动窗口(Sliding Window)的方式。

滑动窗口(Sliding Window)的示例如下：

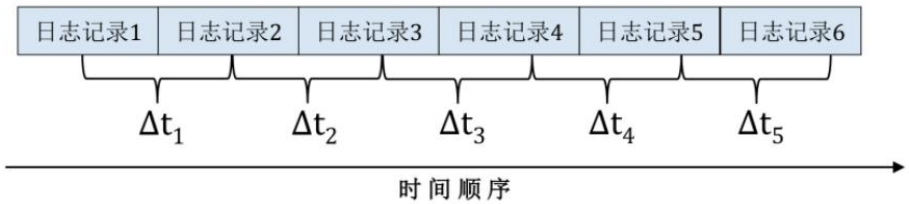

图 3.2.3-1 滑动步长为 1，窗口长度为 2，日志总长为 6 的示例

随后我们通过构建一个基于 LDA 主题模型的日志模版分类模型来将分组完毕后的结构化的日志转化为特征向量，将此特征向量输入到后续的异常检测模块之中。

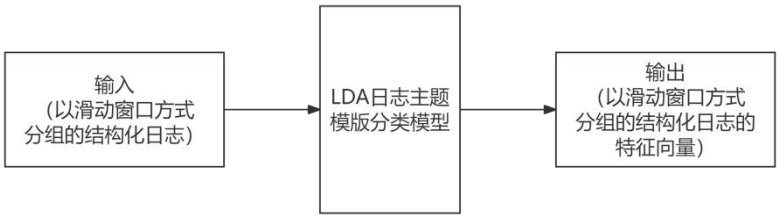

图 3.2.3-2 LDA 日志主题模版分类模型

3.2.4 异常检测[5]

经过特征提取，原始日志数据已经转换为模型可以处理的特征数据，可以输入判别模型进行异常检测。在检测模型的设计上，包括传统机器学习方法和深度学习方法。

通过查阅相关资料我们得知，传统机器学习方法具有硬件依赖性低、可解释性好等特点，但是传统机器学习算法提取高级特征或者全局特征的能力相对有限，特别是日志文本的语义识别、长距离依赖等问题上表现不如深度学习，所以有大量研究将深度学习引入日志异常检测任务。

故我们计划先对传统机器学习检测日志异常的典型算法做基础性的了解，然后学习深度机器学习对日志异常的检测，并付诸实践。

在查阅多篇文献后，我们决定采用 LSTM（Long Short Term Memory）循环神经网络。LSTM 是循环神经网络的一种改进模型，相较于传统的 RNN(Recurrent Neural Network)循环神经网络，LSTM 通过将记忆状态拆分为长期记忆和短期记忆，从而解决了传统 RNN 容易丢失长期记忆信息的缺陷问题。

在我们的项目中，我们选择搭建一个具有 3 个全连接层，每层含有 5 个 LSTM 神经元，每个神经元的输入为日志模版主题的独热码，输出层使用 softmax 函数转化为日志模版主题的概率分布来预测潜在的下一个日志模版主题。

判断是否存在异常的标准：选取输出中前 k 个概率分布最大的向量，如果实际的日志模版主题在这 k 个向量当中，那么就认为不存在异常，否则认为存在异常。

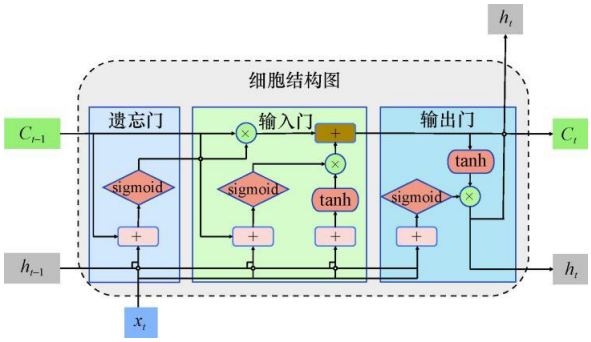

图 3.2.4-1 LSTM 细胞结构

3.3 项目总体架构

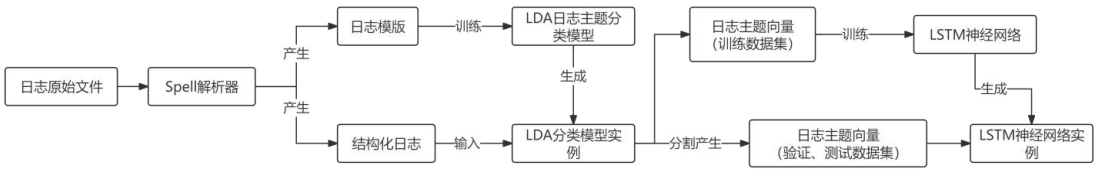

图 3.3-1 项目总体架构

4 结论（成果介绍）

目前，该项目已经完成了中期检查时所完成的全部目标，完成了环境的部署、神经网络的搭建和模型的训练，构建了一个能够使用 LSTM 神经网络模型进行日志异常检测系统，并在此基础上实现了图形化界面，创建了功能合体、代码封装的简单 UI 界面。相比于项目中期，如今的模型的泛化效果更好，训练能力更强，得到的结果也更加准确。

以下是项目的 UI 界面及其运行结果的部分内容：

图 4-1 项目 UI 主界面

图 4-2 训练 LDA 日志分类模型界面

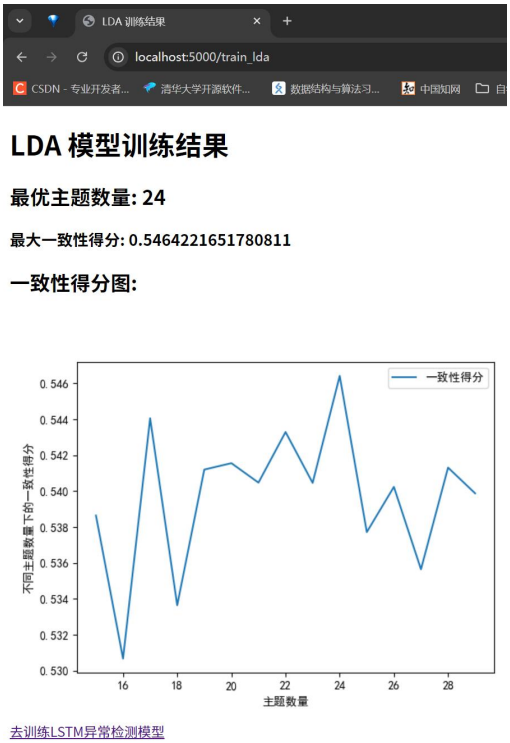

图 4-3 LDA 模型训练结果

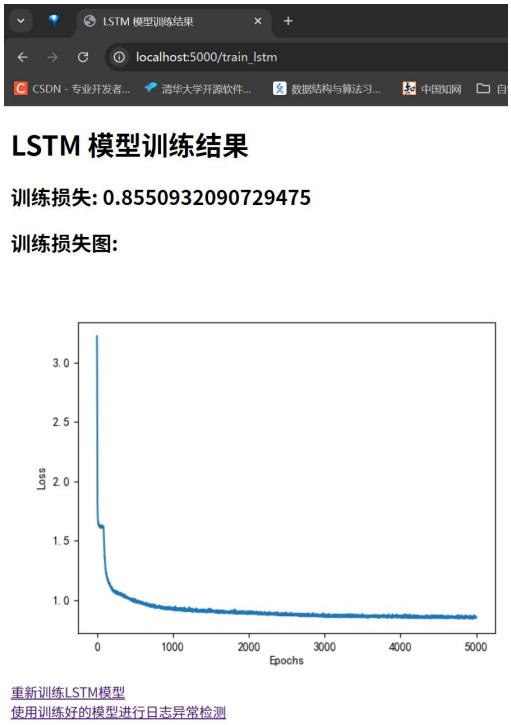

图 4-4 LSTM 模型训练界面

图 4-5 LSTM 模型训练结果

图 4-6 日志推理界面

图 4-7 检测结果

5 经费使用情况

资料费（书名）	用途	金额（元）	合计
《Python 编程 从入门到实践》	Python 入门、基础语法	69.8	174.5
《动手学 深度学习 (Pytorch 版)》	现代深度学习知识	69.8	
《深度学习入门 基于 Python 的理论与实现》	深度学习入门、基础	34.9	

6 问题、体会与收获

在实施项目的过程中，我们对机器学习和深度学习有了一个全新的、更加全面且深度的认知，从线性神经网络到多层感知机再到循环神经网络 CNN 和现代循环神经网络 LSTM、GRU，我们对于整个深度学习的发展历史有了初步的认识。

我们还锻炼了搜集信息和资料的能力，尤其是在如何去实现这个项目的过程中，寻找有关教程和相关文件的能力。通过环境的配置和部署以及解决过程中遇到的问题，让我对计算机系统和命

令行有了更熟练的掌握和使用，提高了借助互联网的帮助解决环境问题的能力。

7 建议

为了进一步提升项目成果的应用价值和社会影响力，我们对本项目提出以下建议：

1.增强数据集多样性：目前本系统的日志异常检测针对 HDFS 日志，支持日志的种类相对单一，我们希望在未来能够获取更多类型的日志作为训练数据集，以获得本系统更加广泛的应用。

2.优化用户界面：目前本系统的用户界面使用的是后端 Flask+前端 HTML 的简单 Web 界面，我们希望在未来可以使用更美观和动态的 Web 界面，实现给用户更好的使用体验。

3.持续技术迭代：目前本系统的核心基于 LSTM 神经网络，一方面 LSTM 发布时间较早，在当下有些过时，另一方面后续也涌现出许多如 transformer 等针对时间序列的更好的更新的神经网络，我们希望能坚持对本系统的持续技术迭代。

8 结束语与致谢

本项目通过结合机器学习，成功构建了一个日志异常检测系统。在项目实施过程中，我们克服了诸多技术难题和挑战。我们相信，随着机器学习的不断发展，必然会为海量日志异常检测这一课题带来新方法和新技术。

在此，我们衷心感谢哈尔滨工业大学本科生院及计算学部的大力支持和悉心指导。感谢项目指导教师王宏志教授和海量数据计算中心学长的专业指导和宝贵建议。同时，我们也要感谢项目团队成员的辛勤付出和紧密合作。没有你们的支持和帮助，我们无法取得今天的成果。我们将继续努力、不断前行，为项目的完善和应用，做出更多努力。

9 参考文献

[1]Wei Xu, Ling Huang, Armando Fox, David Patterson, Michael Jordan. Detecting Large-Scale System Problems by Mining Console Logs, in Proc. of the 22nd ACM Symposium on Operating Systems Principles (SOSP), 2009.

[2]Jieming Zhu, Shilin He, Pinjia He, Jinyang Liu, Michael R. Lyu. Loghub: A Large Collection of System Log Datasets for AI-driven Log Analytics. IEEE International Symposium on Software Reliability Engineering (ISSRE), 2023.

[3][ICSE'19] Jieming Zhu, Shilin He, Jinyang Liu, Pinjia He, Qi Xie, Zibin Zheng, Michael R. Lyu. Tools and Benchmarks for Automated Log Parsing. International Conference on Software Engineering (ICSE), 2019.

[4][DSN'16] Pinjia He, Jieming Zhu, Shilin He, Jian Li, Michael R. Lyu. An Evaluation Study on Log Parsing and Its Use in Log Mining. IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2016.

[5]Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. 2017. DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17). Association for Computing Machinery, New York, NY, USA, 1285–1298. https://doi.org/10.1145/3133956.3134015

六、附件（专利、发表论文及其他成果支撑材料）