Log Anomaly Detection Method based on Hybrid Transformer-BiLSTM Models

Xuedong Ou and Jing Liu
College of Computer Science, Inner Mongolia University, Hohhot, China
1374246122@qq.com, liujing@imu.edu.cn

***Abstract*—Log analysis is quite significant for reliability issues in large cloud data centers. There are noticeable problems in log anomaly detection, such as single feature extraction, unsatisfactory anomaly detection effect. In this paper, we propose a novel log anomaly detection method, which could be divided into two related parts. First, a dataset partitioning method is proposed, named K-fold Sub Hold-out Method (KSHM), which is built on the features of logs to preserve the temporality of training data when sampling. KSHM could enhance the effectiveness of sampling without increasing the number of samples, and change the way the model is trained. Second, an anomaly detection model based on hybrid Transformer-BiLSTM (TFBL) is well constructed, which could extract both temporal and semantic features of logs to serve as a source of features for comprehensive anomaly detection. Experiment results show that TFBL outperforms baseline methods in assessment criteria of accuracy, precision and F1-score, and our log anomaly detection method based on integrated KSHM and TFBL also has better anomaly detection performence.**

***Keywords*—Log anomaly detection; Transformer; Bi-LSTM; Resilient cloud data center**

I. Introduction

Effective log analysis is quite significant for reliability issues in large cloud data centers. Faced with the sheer volume and complexity of logs from cloud data centers, many automatic log anomaly detection methods based on traditional machine learning methods have been proposed [1], [2]. And with wide application of deep learning, various method based on deep learning are used for log anomaly detection [3]–[6]. Although these methods have more powerful learning ability compared to traditional machine learning methods, but they still have certain deficiencies such as single-architecture and the temporal features of the logs is destroyed.

In order to better deal with above problems, we propose a log anomaly detection method based on the features of the log, it consists of two related works. First, a novel K-fold Sub Hold-out Method is proposed, named as KSHM, according to the characteristics of the log. It could preserve the time-series of training data, enhance the validity of training data, and change the way the model is trained. Second, we propose a novel hybrid Transformer-BiLSTM (TFBL) model, which could synergistically extract log temporal features and semantic features.

II. Dataset Partitioning Based on KSHM

For log anomaly detection, previous research often only focuses on the detection model itself and the training method of the model, ignoring the dataset division method. In fact, the data set partitioning method plays a very important role in the whole process. So we introduce KSHM as a dataset partitioning method.

The major process of KSHM is shown in Figure 1, which is composed of four related steps.

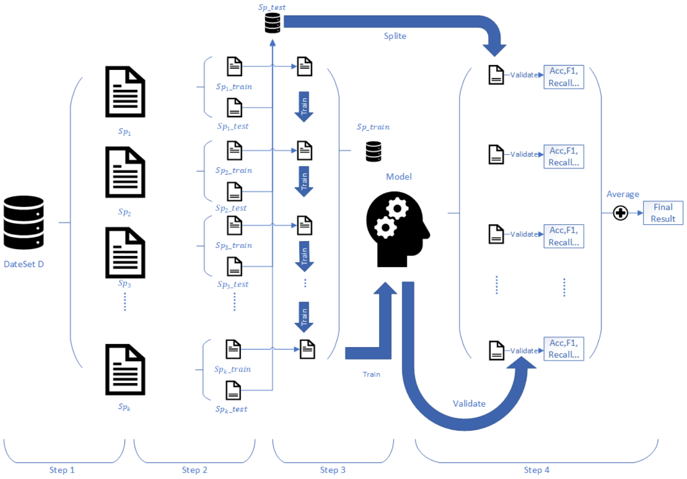

Figure 1. The major process of our K-fold Sub Hold-out Method

- Step1: We need to focus on all parts of the log evenly. The average partition D is $D = \{sp_1, sp_2, sp_3, \ldots, sp_{k-1}, sp_k\}$.
- Step2: We need to determine a partition ratio, here we assume that the training set and testing set are divided in the ratio of 8:2. Then let $sp_k = sp_k\_train + sp_k\_test$. Thus we get a training set $sp\_train = \{sp_1\_train \ldots sp_k\_train\}$ and a testing set $sp\_test = \{sp_1\_test, \ldots sp_k\_test\}$.
- Step3: In the training phase of the model, We use segmented training approach. Each part of the sp_train is fed into the model for T-round training separately in turn. This training way can ensure the training subsets are continuous.
- Step4: In the final model testing stage, we use $\{sp_1\_test \ldots sp_k\_test\}$ in sp_test to test the ability of the model in turn. We need to sum up the results and find an average to obtain the final results.

III. Anomaly detection model based on TFBL

Based on the reasonable partitioning of the dataset, in order to make full use of the high-quality training set brought by

2693-9371/22/$31.00 ©2022 IEEE
DOI 10.1109/QRS-C57518.2022.00123

KSHM, we can then jointly use the TFBL model to effectively extract the features in the training set and make the results of anomaly detection more accurate.

Figure 2 illustrates the architecture of the TFBL model, which consists of the following parts.

- Pre-processing: The main function is to transform each unstructured log message into a parsed log key, and then use the sliding window method to process the log key into a log sequence and vectorize it as the input of the log sequence coder; while keeping the corresponding log message and vectorizing it as the input of the log message coder.
- Log Sequence Coder: This module uses Bi-LSTM to build a feature extractor, and the main function is to extract features from log sequences.
- Log Message Coder: This module uses Transformer to build a feature extractor, whose main function is to extract semantic information from log messages.
- Feature Fusion Classifier: The main function is to accept the feature vectors from the log sequence coder and the log message coder, enhance and optimize them, and give an indication of whether the predicted log entry is anomalous or not.

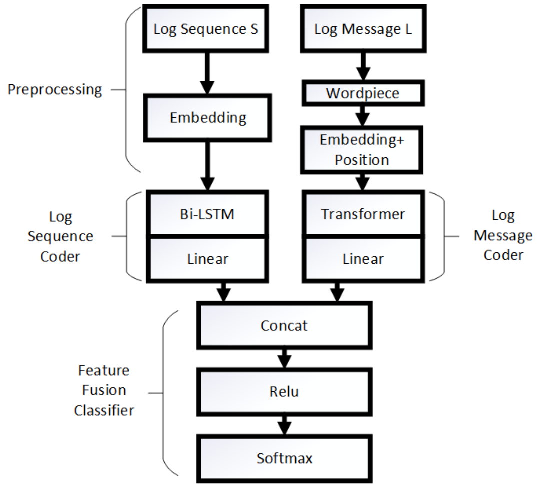

Figure 2. The architecture of our TFBL model

IV. Experiment and Result Analysis

We set up one comprehensive comparative experiments based on the data set division method and the anomaly detection method. We use BGL and HDFS as experimental datasets. The experiment results are well analyzed according to key evaluation metrics.

As shown in Table I, after a comprehensive comparison, we can see that the combination of KSHM and TFBL performs well on both BGL and HDFS. We also find out that the effectiveness of the KSHM for improving the diversity of training samples and enhancing the model's memory ability for anomalous features.

TABLE I
Resutls of comprehensive performance comparison

Method	Dataset	Partition	Accuracy	Precision	Recall	F1
DeepLog	BGL	Hold-out	95.39	95.36	99.928	97.591
		KSHM	99.356	**99.749**	99.589	99.668
	HDFS	Hold-out	97.89	97.904	99.982	98.934
		KSHM	97.912	97.918	99.992	98.942
CT	BGL	Hold-out	95.94	97.476	98.204	97.838
		KSHM	92.554	99.244	92.969	95.565
	HDFS	Hold-out	97.84	97.839	99.997	98.907
		KSHM	74.022	98.356	74.591	81.142
LAMA	BGL	Hold-out	93.51	93.506	**100**	96.644
		KSHM	97.398	97.399	100	98.665
	HDFS	Hold-out	97.83	97.827	100	98.902
		KSHM	97.886	97.886	100	98.929
TFBL	BGL	Hold-out	96.89	99.528	97.137	98.318
		KSHM	**99.504**	99.544	99.949	**99.745**
	HDFS	Hold-out	97.9	97.912	99.982	98.936
		KSHM	97.876	97.912	99.961	98.924

V. Conclusion

In our work, a novel log anomaly detection method is well constructed. First, we propose a dataset partitioning method, KSHM, which is designed based on the time-series, randomness and instability of logs. Then, an anomaly detection model TFBL is developed using a fusion model based on Transformer and Bi-LSTM. It fully utilizes their capabilities in temporal prediction domain to extract temporal features and semantic features in logs for high precision anomaly detection. The comprehensive experiments results demonstrate that our log anomaly detection method based on integrated KSHM and TFBL has better anomaly detection performence.

Acknowledgment

This work was supported in part by the Inner Mongolia Science and Technology Plan Project (No.2020GG0187).

References

[1] S. He, Q. Lin, J. G. Lou, H. Zhang, M. R. Lyu and D. Zhang, “Identifying impactful service system problems via log analysis,” The 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2018, pp. 60-70.

[2] T. Jia, Y. Wu, C. Hou and Y. Li, “LogFlash: Real-time streaming anomaly detection and diagnosis from system logs for large-scale software systems,” The IEEE 32nd International Symposium on Software Reliability Engineering, 2021, pp. 80-90.

[3] M. Du, F. Li, G. Zheng and V. Srikumar, “Deeplog: Anomaly detection and diagnosis from system logs through deep learning,” The 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1285-1298.

[4] X. Zhang, Y. Xu, et al. “Robust log-based anomaly detection on unstable log data,” The 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2019, pp. 807-817.

[5] Y. D. Mei, X. Chen and Y. Z. Sun, “A software system anomaly detection method based on log information and CNN-text,” Chin. J. Computers 2020, 43, pp. 366-380.

[6] Y. Guo, Y. Wen, C. Jiang, Y. Lian and Y. Wan, “Detecting Log Anomalies with Multi-Head Attention (LAMA),” 2021, arXiv:2101.02392, pp. 1-4.