MADDC: Multi-Scale Anomaly Detection, Diagnosis and Correction for Discrete Event Logs

Xiaolei Wang, Lin Yang, Dongyang Li, Linru Ma, Yongzhong He, Junchao Xiao, Jiyuan Liu, Yuexiang Yang

ACSAC, Dec 9th, 2022
Outline

1. Introduction
2. System Design
3. Evaluation
Discrete Event Log Anomaly Detection--An Illustrative Example

**Step 1: Log Event Key Composition**
- tcp,FIN,ftp-data,0,0,0,0,0,0,0,0,0 -- 5
- tcp,FIN,-,0,0,0,0,0,0,6 -- 9
- tcp,FIN,ssh,0,0,0,0,0,0,1 -- 11
- tcp,FIN,-,0,0,0,0,0,3,0 -- 22
- tcp,FIN,-,0,0,0,0,0,19,0 -- 26

**Step 2: Log Event Sequence Aggregation**

**Step 3: Directly Follows Graph (DFG) Construction**

**Abnormal:** 22-5-5-5-26-26-26-11-9-11
Recent Trends - Deep Learning based Anomaly Detection

**DeepLog**

Input: $h$ recent log event keys up to $e_{t-1}$

$w = \{e_{t-h}, \ldots, e_{t-2}, e_{t-1}\}$

Output: conditional prediction probability of next event key given the recent sequence

Check Whether Real Next Log Key is in Top-N Predictions

Top N Candidates (N=3)

**DabLog**

Input: \{\(e_{t-h}, \ldots, e_{t-2}, e_{t-1}\)\}

Output: conditional reconstruction probability of each key at each time slot

Check Whether Real Original Log Key is in Top-N Predictions

Top N Candidates (N=3)

Improved a lot the accuracy of anomaly detection for discrete event logs. But still is not enough!!!

Two known LSTM based works, **DeepLog** (CCS 2017), **DabLog** (Asiaccs 2021)
Limitations of Existing Representative Deep Learning based Methods

To Fulfil Practical Anomaly Detection for Discrete Event Log:

• Accuracy of Anomaly Detection Needs Further Improvement.
  o Ability to handle event logs with complex temporal correlation.
    ▪ DeepLog’s next event prediction tends to be more frequency based.
  o Alleviate model overfitting.
    ▪ DabLog fails to characterize abnormal regions in the latent space.

• Anomaly Diagnosis Needs More Attention to Improve Interpretability. **Our Main Focus**
  o Accurate Abnormal Deviation Identification. *Why abnormal?*
  o How Normal Pattern Should Behave. *How make correction?*
Motivation for Anomaly Diagnosis (that DabLog’s sliding window based reconstruction cannot do)

The abnormal event 13, it is expected to be reconstructed as 5 in window 3 but as 22 in window 4

The tenth event 9, detected as normal in sliding window 3, but detected as abnormal and reconstructed as 26 in window 4.

Cannot provide consistent diagnosis for potential correction

Fail to Identify Abnormal Deviation Accurately

The above inconsistency regards to diagnosis is because each sliding window provides a limited and variable view.

Number in black box is Top-N (here N=1) candidate event with the highest reconstruction probability at each time slot.
Our Solutions

- **Key Insights**

  - **Local** sliding-window (**small-scale** view) based anomaly detection can provide better precision, but not suitable for anomaly diagnosis.

  - **Global** workflow (**large-scale** view) based whole sequence alignment can accurately uncover how an anomaly deviates from the “normal” pattern, thus facilitating anomaly diagnosis, understanding and correction.

  Why not combine them?

Our sequence alignment based anomaly diagnosis results of the motivating example, as follows:

```
```
Outline

1. Introduction

2. System Design

3. Evaluation
Overview

- **Proposed System Prototype-MADDC**
  - Three major separate components, namely *log preprocess, local anomaly detection, global anomaly diagnosis and correction*.
  - Combine **LSTM-based** Variational AutoEncoder with **Process Mining Techniques** (i.e., process discovery and conformance checking).

- **Overall Process**

  ![Diagram](image-url)

  - Log Sequences → Preprocess
  - Process Mining based Workflow Construction
  - Detect whether a log sequence $L$ is abnormal in a sliding fashion
  - Global alignment based anomaly diagnosis on whole $L$
Several Key Definitions

▪ **Event Keys**
  - Log entries are parsed using templates and represented by a **number**.

▪ **Event Sequence and Subsequence**
  - A log sequence is transformed into an event sequence using event keys.

▪ **Process**
  - During process mining, we treat each **event sequence** as a **process**.

▪ **Workflow**
  - A workflow is defined at the higher level to characterize a **group of similar processes**, namely closely related cases which may execute similar tasks.
LSTM-VAE based on Local Anomaly Detection

- **Subsequence Reconstruction using LSTM-VAE (Variational Autoencoder)**

  Event Subsequence
  \[ \{e_1, e_2, \ldots, e_n\} \]
  \( \rightarrow \) Encoder-Decoder
  \( \rightarrow \) \( \{e'_1, e'_2, \ldots, e'_n\} \)

  Check event reconstruction at each time slot

- **Double-Check Anomaly Criterions**

  Example: \([11 \ 13 \ 11 \ 28 \ldots]\)

  1) Top-K Rank-based Criterion

     Original event at time slot 4 is 28.

     Check whether is included √

     The Top-3 events sorted by reconstruction probability at time slot 4 is \([11, 28, 4]\)

  2) Probability Threshold-based Criterion

     Calculates the occurrence (reconstructed) probability \(\theta_p\) of reconstructed event \(e_j\) being same as 28

     \(\theta_p = 0.02859\) X

     Check whether the occurrence (reconstructed) probability \(\theta_p\) at slot 4 exceeds a predefined threshold \(\theta\).
The main principle behind our anomaly diagnosis is to uncover critical differences by comparing the detected abnormal sequence with a collection of similar “normal” ones.

**Challenge 1:** Unclear “normal” sequence pattern for anomaly diagnosis

**Challenge 2:** Accurate Abnormal Deviation Identification, avoiding time consuming ‘one-to-many’ sequence comparison
Alignment based Anomaly Diagnosis—An Example

- **Goal of alignment**
  - Map observed behavior (i.e., event sequence) from logs onto modelled behavior (i.e., workflow model) to derive deviations and conformance on event level.

- **Trace based alignment**
  \[ \gamma_2 = \begin{array}{cccc}
  a & b & c & d & e \\
  a & b & c & d & e \\
  \end{array} \]

  - The ‘›’ indicates that either the log could not make the same step as in the model or vice versa, considered as anomaly.
  - Anomaly behaved as **two kinds of** asynchronous moves:
    - Extra events -- cases where the log events makes move, but unallowed by the model.
    - Missing events -- cases where there are moves in model but not in log events.

- **Alignment based anomaly correction**
  - Based on alignment results, try to correct the anomaly by removing extra events and adding missing events at the corresponding position.
Outline

1. Introduction

2. System Design

3. Evaluation
Evaluation

- Research Questions:

  ➢ Question 1: How better is MADDC in anomaly detection when compared with representative reconstruction-based and prediction-based baseline models?

  ➢ Question 2: As a key factor to provide reliable alignment based anomaly diagnosis, what quality are workflow models built on clustered sequences?

  ➢ Question 3: How effective does alignment based anomaly diagnosis facilitate the anomaly understanding and interpretation?
Experimental Setup

- **Datasets and Models:**
  - **Datasets:** UNSWNB (intrusion detection traffic logs), HDFS system logs.
  - **Models:** DeepLog(CCS 2017), DabLog(AsiaCCS 2021)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Normal Train</th>
<th>Normal Test</th>
<th>Abnormal Test</th>
<th>Number of Event Keys</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sequences</td>
<td>Subsequences</td>
<td>Sequences</td>
<td>Subsequences</td>
</tr>
<tr>
<td>HDFS-1</td>
<td>4855</td>
<td>61,140</td>
<td>553,366</td>
<td>6,918,652</td>
</tr>
<tr>
<td>HDFS-2</td>
<td>194,115</td>
<td>2,425,217</td>
<td>194,066</td>
<td>2,428,025</td>
</tr>
<tr>
<td>UNSWNB</td>
<td>9900</td>
<td>1,843,301</td>
<td>9900</td>
<td>1,853,201</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Both HDFS-1 and HDFS-2 are generated from original HDFS Dataset.
2. HDFS-1 is the same dataset used in DeepLog.
3. HDFS-2 is generated using same method as in DabLog.
Accuracy of Anomaly Detection

- We have reproduced DeepLog on HDFS-1 with very similar performance.

Table 2: Anomaly Detection Results of Models when $\pi=0.1$, $\theta=0.1$

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>FP</th>
<th>FN</th>
<th>P $^3$</th>
<th>R $^4$</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDFS-2</td>
<td>DeepLog</td>
<td>9927</td>
<td>5838</td>
<td>44.58%</td>
<td>57.76%</td>
<td>50.32%</td>
</tr>
<tr>
<td></td>
<td>DabLog</td>
<td>267</td>
<td>2777</td>
<td>97.65%</td>
<td>80.00%</td>
<td>87.95%</td>
</tr>
<tr>
<td></td>
<td>MADDC</td>
<td>335</td>
<td>895</td>
<td>97.49%</td>
<td>93.55%</td>
<td>95.48%</td>
</tr>
<tr>
<td>UNSWNB</td>
<td>DeepLog</td>
<td>2996</td>
<td>196</td>
<td>40.61%</td>
<td>91.27%</td>
<td>56.21%</td>
</tr>
<tr>
<td></td>
<td>DabLog</td>
<td>978</td>
<td>207</td>
<td>67.57%</td>
<td>90.78%</td>
<td>77.48%</td>
</tr>
<tr>
<td></td>
<td>MADDC</td>
<td>27</td>
<td>110</td>
<td>98.75%</td>
<td>95.10%</td>
<td>96.89%</td>
</tr>
</tbody>
</table>

$^3$ Precision Rate, $^4$ Recall Rate.

- Detailed parameter analysis shows the MADDC’s stable performance on different dataset with varying parameters.

- $\pi = N$ / Number of event keys, rank based parameter
- $\theta$, probability threshold-based parameter
- FP: false positives identified through manual check.
  - Due to rare patterns
- FN: false negatives identified through manual check.
  - Subsequences pattern that are abundant in the training set
Case Study and Advantages on Accurate Anomaly Detection

- **Double-check** based anomaly critic could make full use of their respective advantages.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ID</th>
<th>Anomaly Subsequence</th>
<th>MADDC’s Reconstruction &amp; Probability</th>
<th>Dablog’s Reconstruction &amp; Probability</th>
</tr>
</thead>
</table>

- **Different Event Reconstruction Probability** from MADDC and Dablog (0.07836<<1.0): Due to VAE’s probabilistic modeling of MADDC, the latent distribution of abnormal data have greater variance.
Effectiveness of Alignment based Anomaly Diagnosis

- **Our Alignment is** **consistent** and **accurate** for anomaly diagnosis.
  - Extra events: 7, 17; Missing events: 16.

- DeepLog’s next event prediction is **consistent** but **inaccurate** for anomaly diagnosis.
  - Ten detected abnormal events, eight of which are manually confirmed as FPs.

- Dablog’s subsequence reconstruction is **inconsistent** and **inaccurate** for anomaly diagnosis.
  - Event 18 is not abnormal in window 2, but is in next window.
  - Event 18 again, its Top-1’s reconstructed event is 5 in window 3 but becomes 3 in the next window.

---

Table 5: Case Study 1-A HDFS-2 Anomaly Diagnosis

<table>
<thead>
<tr>
<th>Alignment</th>
<th>5 5 22 5 11 9 11 9 11 9 25 26 26 5 18 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17 16 3 23 23 21 21 21</td>
</tr>
<tr>
<td>Sliding</td>
<td>5 5 22 5 11 9 11 9 11 9 25 26 26 26-23 5 18</td>
</tr>
<tr>
<td>Fashion</td>
<td>9 11 9 25 26 26 5 18 7-16</td>
</tr>
<tr>
<td>Prediction</td>
<td>5 5 22 5 11 9 11 9 11 9 25 26 26 26-23 5 18</td>
</tr>
</tbody>
</table>

The circled number 4 refers to the Top-1 predicted or reconstructed event.
What We Have Not Talked About

- Unsupervised Characterization of “Normal” Sequence Pattern
  - Why Event Sequences Clustering?
  - What is Process Discovery based Workflow Construction?
- Quality Evaluation of Workflow Model Construction in Unsupervised Manner
- Accuracy Comparison of Alignment based Anomaly Diagnosis.
- Limitations and Future Work.

Please Read Our Paper!
Thank You!

Xiaolei Wang
xiaoleiwang@nudt.edu.cn
https://github.com/040840308/MADDC/tree/master