Device-Agnostic Log Anomaly Classification with Partial Labels

Weibin Meng, Ying Liu, Shenglin Zhang, Dan Pei
Hui Dong, Lei Song, Xulong Luo
Motivation

Architecture of Datacenter Networks

Inter-DC Network

ToR Switch

Server

Aggregation Switch

Access Router

Core Router

IDPS

Firewall

VPN

Load balancer

IDPS

Firewall

VPN

Load balancer

L3

L2

Core

Load balancer

ToR Switch
Motivation

- Traditional anomaly detection methods usually monitor KPI curves.
  - KPI need network operators select manually.
  - KPI methods can only find anomalous behaviors.

- Logs describe some events that KPI curve can’t, such as the root cause.
- Logs are most valuable data sources for device management.

```
SYSLOG/6/SYSLOG_RESTART: System restarted -- H3C Comware Software.

DEV/Z/FAN STATE CHANGE TO FAILURE: Trap 1.3.6.1.4.1.2011.2.23.1.12.1.6(fanfailure): fan ID is 1

P01 OUT_SWITCH 192.168.201.218 2016 %%10DEVM/1/FAN STATE CHANGES TO FAILURE(t): Trap
1.3.6.1.4.1.2011.2.23.1.12.1.6: fan ID is 1

DEV/SYSTEM_REBOOT: System is rebooting now.
```
# Device logs

- **Examples of device(switch) log**:

<table>
<thead>
<tr>
<th>Switch ID</th>
<th>Timestamp</th>
<th>Message Type</th>
<th>Detailed Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switch 1</td>
<td>Jun 12 19:03:27 2017</td>
<td>SIF</td>
<td><strong>Interface</strong> te-1/1/59, changed state to down</td>
</tr>
<tr>
<td>Switch 2</td>
<td>Jun 13 20:22:03 2017</td>
<td>-</td>
<td>Vlan-interface vlan22, changed state to down</td>
</tr>
<tr>
<td>Switch 1</td>
<td>Jun 13 20:22:03 2017</td>
<td>SIF</td>
<td><strong>Interface</strong> te-1/1/17, changed state to up</td>
</tr>
<tr>
<td>Switch 18</td>
<td>Jun 18 05:21:03 2017</td>
<td>SIF</td>
<td><strong>Interface</strong> te-1/1/19, changed state to up</td>
</tr>
<tr>
<td>Switch 22</td>
<td>Jun 15 13:46:43 2017</td>
<td>OSPF</td>
<td>Neighbour vlan23, changed state from Exchange to Loading</td>
</tr>
</tbody>
</table>

Detailed Messages are **Semi-structured natural languages** provided by device developers.

Message types are ambiguous for accurate classification.
Drawbacks in Regular Expression

- **Regular Expression** is the popular technique for anomalous log classification.

- **Drawbacks:**
  - Low generality
  - Labor intensity
  - ...

Configure anomalous regular expressions

Syslog → Match

- Yes → Type 1 → RE for Manufacturer A
- No → Type n
- Ignore

Manufacturer B logs
Problem Definitions

Anomalous log detection  Anomalous log classification

First Category of Anomalous Logs

Second Category of Anomalous Logs

Anomalous Logs

Healthy Logs

Device Logs
Challenges

• **Device-agnostic vocabulary**
  - Device logs are type-specific and manufacturer-specific.
  - It is hard to fit one classification model for all different device types.

• **Partial labels**
  - Network operators only label partial anomalous logs they encountered.
  - Difficult to train a traditional classification model.
LogClass Design Overview

1. Log Preprocessing
2. Feature vector
3. Anomaly detection
4. Anomaly classification
Text feature vector

The universal method to construct a text feature vector is the bag-of-words model.

logs:

<table>
<thead>
<tr>
<th>Interface</th>
<th>vlan22</th>
<th>changed</th>
<th>state</th>
<th>to</th>
<th>down</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface</td>
<td>changed</td>
<td>state</td>
<td>to</td>
<td>down</td>
<td></td>
</tr>
<tr>
<td>VlanInterface</td>
<td>changed</td>
<td>state</td>
<td>to</td>
<td>up</td>
<td></td>
</tr>
<tr>
<td>Neighbour</td>
<td>changed</td>
<td>state</td>
<td>from</td>
<td>to</td>
<td>Loading</td>
</tr>
</tbody>
</table>

bag-of-words vectors:

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Interface</th>
<th>changed</th>
<th>state</th>
<th>to</th>
<th>down</th>
<th>VlanInterface</th>
<th>Neighbour</th>
<th>from</th>
<th>Exchange</th>
<th>Loading</th>
<th>up</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_1)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(L_2)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(L_3)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Assign weighting values to each component in vectors. (e.g., TF-IDF)
PU Learning

- Different from tradition classification.
  - In our scenario, labelling all existing anomalous logs is not natural.
- PU Learning input:
  - Positive set $P$ (Anomalous logs)
  - Unlabeled set $U$ (Unlabeled logs)

(Gang Niu et al. NIPS’16)

2018/6/23
Evaluation

Dataset

• Real-world Switch logs
• 58 switches types
• Two-week period
• 1,758,456 anomalous logs
• 16,702,547 unlabeled logs

Benchmark methods

• Labeled-LDA
• Regular Expression
Evaluation on PU Learning

Sampled anomalous logs randomly cross all switch types and assumed they have no labels.

PU Learning classifier is more stable than traditional classifier.
Evaluation on Anomalous Log Classification

<table>
<thead>
<tr>
<th>Methods</th>
<th>Macro-F1</th>
<th>Micro-F1</th>
<th>Training Time(s)</th>
<th>Classifying Time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogClass</td>
<td>95.32%</td>
<td>99.74%</td>
<td>247.73</td>
<td>4.836</td>
</tr>
<tr>
<td>L-LDA</td>
<td>89.68%</td>
<td>93.53%</td>
<td>4436.4</td>
<td>28.59</td>
</tr>
<tr>
<td>RE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>419.47</td>
</tr>
</tbody>
</table>

LogClass is more accurate.

The overheads of L-LDA and RE are larger than LogClass.
Conclusion

**Challenges**

- Device-Agnostic vocabulary
- Partial anomalous logs have labels

**LogClass**

- PU learning
- Simple NLP techniques

**Evaluation**

- Real-world switch logs.
Thank you!

mwb16@mails.tsinghua.edu.cn