ANM 2018 Fall Assignments and Project

NENGWEN ZHAO

ZNW17@MAILS.TSINGHUA.EDU.CN

Overview

- Assignment #1: Data Preprocessing and Visualization (10%)
- Assignment #2: Log Analysis for Anomaly Detection (20%)
- Project: Time Series Anomaly Detection Algorithm

Competition (60%)

Each student finishes the assignment alone and a team of 2-3 students finish the project together.

What you can learn

- At least one programming language (Python is recommended).
- At least one data visualization tool, such as PowerBI, Matlab, GNUPlot, and matplotlib.
- Some background regarding AIOps (Artificial Intelligence for IT Operations).
- Some machine learning tools such as Scikit-learn, PyTorch, TensorFlow, Keras.
- Using Google to search the things you do not know (important!)

Assignment #1-- Data Preprocessing and Visualization

- 2 weeks of search logs from a global top search engine
- Basic statistics: coding to count the data in different ways, e.g., how many queries are served per minute. You can also use Power BI do it.
- Visualization: plot figures to show the data (e.g., line chart, histogram, and CDF). You can use Power BI, matplotlib, gnuplot, matlab, etc. to do it.

```
Timestamp,#Images,UA,Ad,ISP,Province,PageType,Tnet,Tserver,Tbrowser,Tother,SRT
1411315200,0,Chrome,noAD,CRTC,Heilongjiang,sync,1495.0,443.14,73.0,0.0,2011.14
1411315200,13,Chrome,noAD,CHINANET,Guangdong,async,200.0,80.0,47.0,359.0,686.0
1411315200,24,Chrome,noAD,UNICOM,Hunan,async,120.0,450.0,26.0,191.0,787.0
1411315200,4,Safari,noAD,OTHER,Guangdong,async,272.0,86.0,234.0,219.0,811.0
```

Assignment #1-- Data Preprocessing and Visualization

- Calculate the average SRT of every 10 minutes, and plot the SRT with a line chart (x axis for date time and y axis for the average SRT).
- Calculate the average of each SRT component of every 10 minute, and plot the four SRT components together with a stacked area chart (x axis for date time and y axis for time) and also a 100% stacked area chart (y axis for the percentage).
- Plot the CDF (Cumulative distribution function) chart of SRT.
- Plot the CDF chart of #Images.
- Count the number of queries (also called page views or PVs) of each minute, and plot the minute-level PVs with a line chart (x axis for date time and y axis for the PVs).
- Count the PVs of each province, and plot it with a histogram chart (x axis for province and y axis for PVs).
- Count the PVs of each UA, and plot it with a pie chart (show the percentages in the chart).
- What are the differences among those charts (How to decide which one to use)
- Describe your experience or findings in doing those jobs. For example, experience of processing the data, observations from the charts, characteristics of the data, potential explanations, and any interesting things you would like to mention.

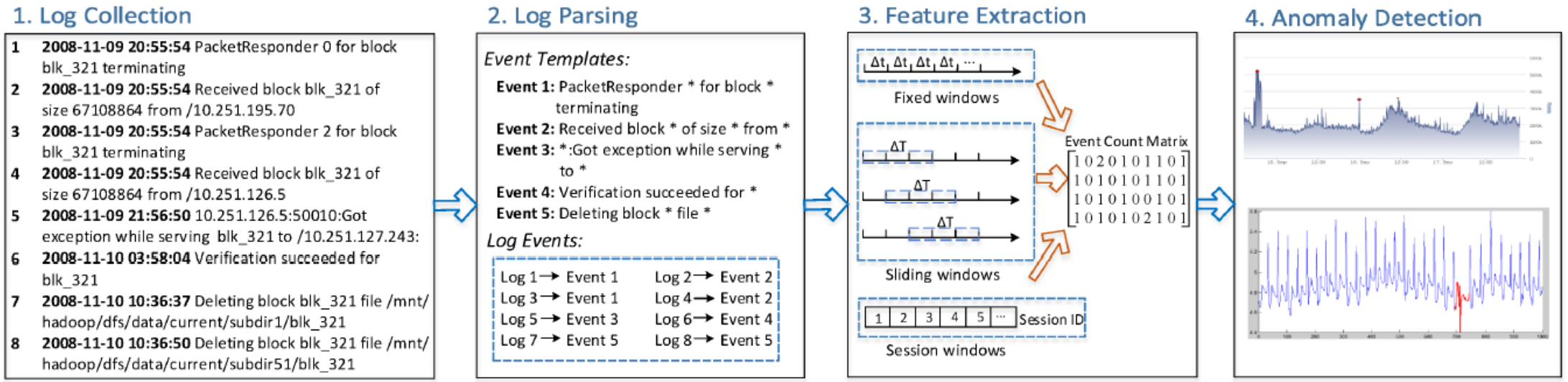
Assignment #2– Log Analysis for Anomaly Detection

- Logs are the main data source for system anomaly detection.
- Logs are routinely generated by systems (e.g., 24 x 7 basis).
- Logs record detailed runtime information, e.g., timestamp, state, IP address.

1	**2008-11-09 20:55:54** PacketResponder 0 for block blk_321 terminating
2	**2008-11-09 20:55:54** Received block blk_321 of size 67108864 from /10.251.195.70
3	**2008-11-09 20:55:54** PacketResponder 2 for block blk_321 terminating
4	**2008-11-09 20:55:54** Received block blk_321 of size 67108864 from /10.251.126.5
5	**2008-11-09 21:56:50** 10.251.126.5:50010:Got exception while serving blk_321 to /10.251.127.243:
6	**2008-11-10 03:58:04** Verification succeeded for blk_321
7	**2008-11-10 10:36:37** Deleting block blk_321 file /mnt/hadoop/dfs/data/current/subdir1/blk_321
8	**2008-11-10 10:36:50** Deleting block blk_321 file /mnt/hadoop/dfs/data/current/subdir51/blk_321

Assignment #2– Log Analysis for Anomaly Detection

Popular Framework of log anomaly detection



Current log parsing methods:

- **LogSig (CIKM'11)**
- **LKE (ICDM'09)**
- **IPLoM (KDD'09)**
- **SLCT (IPOM'03)**

Current anomaly detection methods:

- **Log Clustering (ICSE'17)**
- **PCA (SOSP'09)**
- **Invariants Mining (ATC'10)**

Assignment #2– Log Analysis for Anomaly Detection

Part 1: compare current log parsing methods

Code:

- https://github.com/logpai/logparser
- Algorithms: LogSig, IPLoM, SLCT, LKE

Data:

- https://github.com/logpai/logparser/tree/master/data
- Five datasets (BGL, HDFS, HPC, Proxifier, Zookeeper).

Requirement:

1. Run four algorithms (LogSig, IPLoM, SLCT, LKE).
2. Compare the running time, F-score, RandIndex respectively when four algorithms parse five datasets.

Assignment #2– Log Analysis for Anomaly Detection

Part 2: compare anomaly detection methods

Code:

- https://github.com/logpai/loglizer
- Algorithms: Invariants Mining, PCA and Log Clustering

Data:

- HDFS logs with labels (1.5G)

Requirement:

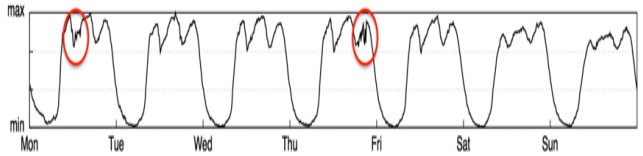
- Choose a log parsing algorithm to change HDFS logs into template sequence, then run three different anomaly detection methods.
- Display precision, recall, F-score, and running time
- Run *invariants mining* and display three invariants.

Project -- Anomaly Detection Competition

Anomaly detection: binary classification problem

Dataset: 27 labeled KPIs from five large Internet companies

- training set: 50%, with label used to train your algorithm
- testing set: 50%, without label used to test your algorithm



KPI ID	timestamp	value	label
A	1411315200	90.75	0
A	1411315260	96.78	1
…	…	…	

Training set

KPI ID	timestamp	value
A	1411423000	83.2
A	1411423060	91.4
…	…	…

Testing set

KPI ID	timestamp	predict
A	1411423000	0
A	1411423060	1
…	…	

Submitted fils

Project -- Anomaly Detection Competition

Requirements:

- Design a generic anomaly detection algorithm and submit your result on testing set in the website. The website will give a rank list of F-score like Kaggle. (Leaderboard: 40%)
- Submit runnable codes and a report about all details of your algorithm, including data preprocessing (normalization? fill missing? etc.), algorithm implementation, parameter setting... (10%)
- Give a presentation. (10%)

Project -- Anomaly Detection Competition

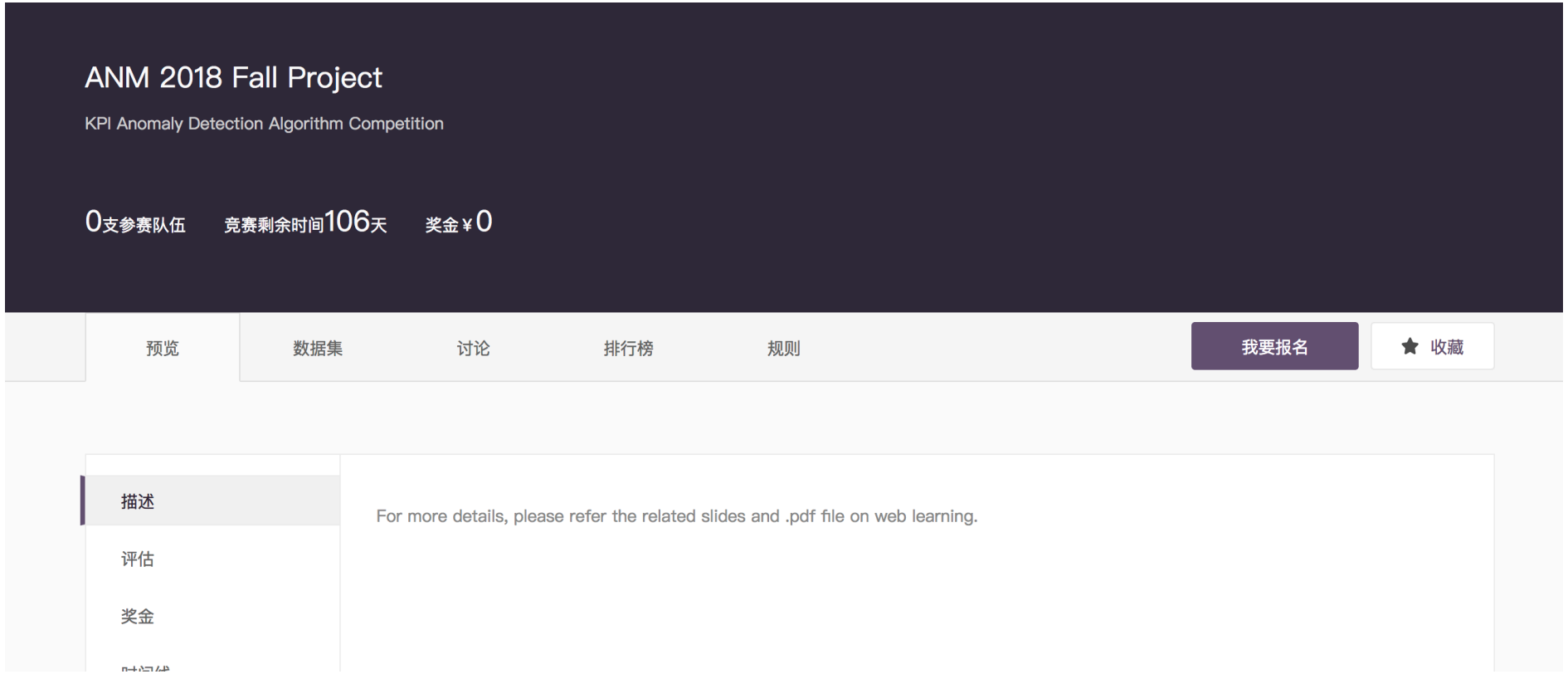
Leaderboard Scoring rule:

- The first place with best F-score will get 40 points.
- Other teams: score = $\frac{your\ F-score}{best\ F-score} \times 40$

  For example, best F-score = 0.8, the F-score of your team is 0.6, you will get

  score = $\frac{0.6}{0.8} \times 40$=30 points

KPI Anomaly Detection Competition Website

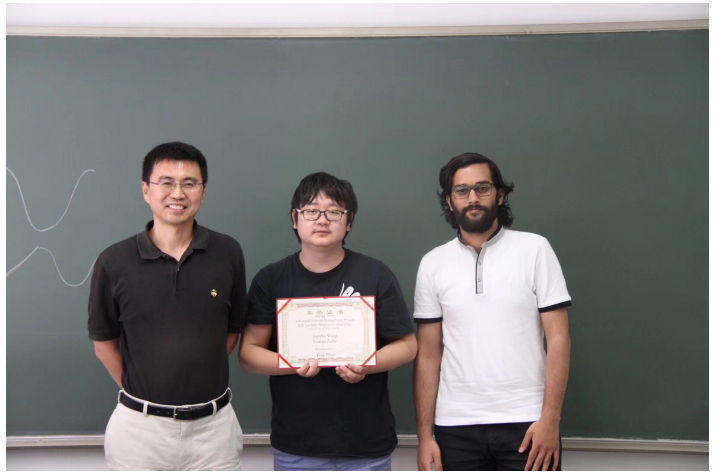


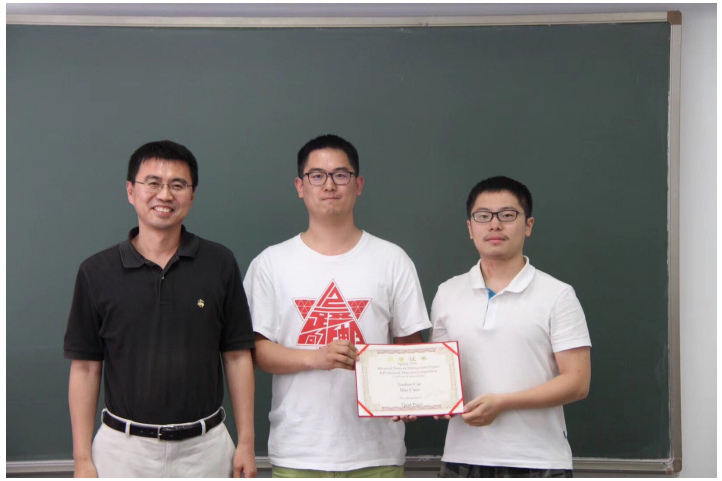
http://iops.ai/competition_detail/?competition_id=7&flag=1

Review: ANM 2018 Spring

队伍排名	队伍名字	队伍分数
1	lovelycute	0.653903353512
2	lele	0.653474872519
3	hyoga	0.634830632077
4	邪王真眼	0.631946850661
5	robwanders	0.582175714587
6	天蚕	0.523525572564
7	Yolanda	0.501922480202
8	A10	0.480756733338
9	Forec_	0.444796252521
10	Hunder	0.438023751992

Review: ANM 2018 Spring





Thank you!

Please get the data files from TA.