Online and adaptive detection of web attacks

Motivation

- A web attacks is one of major threat in current computer networks
  - With over 70% of attacks now carried out over the web application level
- Online detection
  - Unsupervised: no need of labeled data
- Adaptive detection
  - Deal with concept drift problem

Data

- Http log data from INRIA Sopia
  - Original size: 561M
  - N. of request: 1,449,379
  - Duration: 3 days and 2 hours 10 mins
- Data filtering
  - Filtered the robot
  - Filtered most of static request
    - File htm, jpg, gif, pdf, doc…
  - Size after filtering:
    - N. of request: 60,334
      - Only remain 4.16% of the original requests

Data Preprocessing

- Original data form

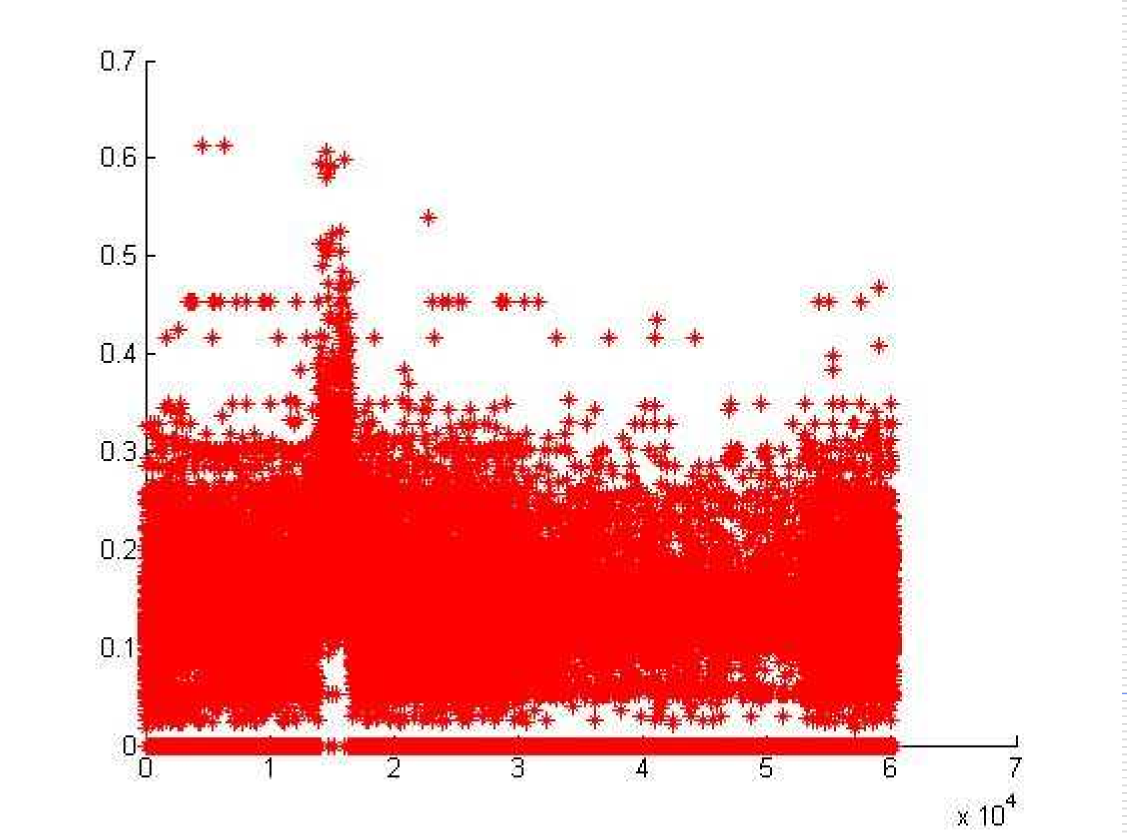
```
salmacis.inria.fr - - [10/May/2007:18:27:32 +0200] "GET /cgi-bin/db4web_c/dbdirname//etc/passwd HTTP/1.0" 404 4856 "-" "Mozilla/4.75 (Nikto/1.36 )"
```

- Computer the character distribution of the request path source
  - Only computer the distribution of ASCII 33-127
  - Each request is thus represented by a vector with 95 dimensions
  - Classification is based on the vectors

Classification

Anomaly detection

- Select the first 200 requests as references (base)
- Compute distances between each coming request and all the first 200 requests
- Select the minimal distance as the anomaly index



Classification

- Change detection
  - Page-Hinkley change-point detection
  - Upgrade the reference if a change point is found
- Work in progress
  - Improve the data preprocessing methods
    - Frequency weights of the character distribution
  - Upgrade the models for incremental learning
  - Better methods for unsupervised learning