UNIVERSITÉ DE STRASBOURG

UFR Mathématique Informatique Strasbourg

Intelligence artificielle – Projet 'détection d'anomalie'

Licence 3 Informatique, semestre 6
Année 2014-2015
TPs: Pierre Parrend

Objectifs du projet

L'objectif du projet est de développer un outil simple de détection d'anomalies de sécurité dans des logs informatiques.

Les livrables du projet sont :

- Une application fonctionnelle de détection d'anomalies dans une base de logs
- Une présentation, y compris une démonstration

Le dataset standard KDDCup99 sera utilisé.

Déroulement

- Date de présentation du projet : 14/4 (groupe du mardi), 24/4 (groupe du vendredi)
- Travail par groupes de 3 ou 4
- Présentation
  - Slides + démonstration (après les slides !)
  - 10 min par groupe
- Rendu
  - Le code sera mis à disposition la veille de la présentation via Google Drive ou équivalent (accès en lecture sans invitation nécessaire)

Cahier des charges

L'application développée

- est basée de préférence sur Python/Django
- inclue 1 interface graphique, utilisant de préférence la bibliothèque. D3JS : http://d3js.org/. L'interface graphique pourra être composée de 2 écrans :
  - 1 écran de sélection du dataset et de paramétrage de l'analyse (K et N tels que définis ci-dessous ; choix des champs utilisés pour la classification)
  - 1 écran de détection d'anomalies affichant les classes de données de manière graphique en 2 ou 3 dimensions. Un code couleur permettra de visualiser les classes, ainsi que les N% de comportements en marge des classes de comportement.
  - Optionnel : chaque log est représenté, et est accessible directement par l'écran de détection d'anomalie
- inclue 1 bibliothèque de classification et d'identification des anomalies

    - La classification se fait en mode non supervisée, sur la base de fichiers de logs
    - 1 fonction permet l'extraction de K classes de comportement (K en paramètre)
    - 1 fonction permet d'extraction des N% de comportements les plus en marge de chaque classe, dans un ordre de distance décroissante pour la classe (N/100 en paramètre)
    - L'algorithme de classification utilisé pourra être K-Means
- incluant des tests unitaires pyunit pour cette bibliothèque

L'utilisation de langages et d'outil alternatifs est acceptée.

Le choix des champs de données du dataset à utiliser comme référence pour la détection d'anomalies fait partie du travail à réaliser par le groupe, ainsi que le choix des paramètres K et N.

Dataset

Vous utiliserez le dataset de référence de logs systèmes KDDCup99, disponible ici :
http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
En particulier, les fichiers :

- kddcup.names
- kddcup.data_10_percent

vous serons utiles.
Le fichier :
http://www.takakura.com/Kyoto_data/BenchmarkData-Description-v3.pdf
contient des informations complémentaires concernant les champs de données du dataset KDDCup99.