Intelligence artificielle – Projet 'détection de spam'

Licence 3 Informatique, semestre 6
TPs: Pierre Parrend

Objectifs du projet

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

Les livrables du projet sont :

- Une application fonctionnelle de détection de spams dans une base de données de traces
- Une présentation, y compris une démonstration

Le dataset 'SpamDatabase', mis à disposition par les Hewlett-Packard Labs via le Machine Learning repository de l'UC Irvine sera utilisé.

Déroulement

- Date de présentation du projet : 5/4 (groupe du mardi), 22/4 (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)

2 projets au choix

Sur la base du dataset proposé, vous choisirez l'une des deux approches suivantes :

- Utilisation de SVM (Support Vector Machine) pour effectuer la classification des données (sur 4000 lignes) puis tester l'outil (sur 600 lignes)
  - Ecran 1 : choix du dataset et des champs de données
  - Ecran 2 : affichage simple des résultats de la classification
  - Ecran 3 : affichage graphique des résultats du test, en 2D ou 3D (rouge : spam ; bleu : non spam)
- Utilisation de K-Means pour réaliser un clustering des données, et visualiser la pertinence des clusters.
  - Ecran 1 : sélection du dataset et de paramétrage de l'analyse (K et N tels que définis ci-dessous ; choix des champs utilisés pour la classification)

- Ecran 2 : affichage des propriétés statistiques des différents champs normalisés sur un espace [0 ;1] (min, max, moyenne, écart type), permettant également le choix des 2 (pour affichage en 2D) ou 3 champs (pour affichage en 3D) utilisés pour la classification
- Ecran 3 : détection de spam affichant les classes de données de manière graphique en 2 ou 3 dimensions. Chaque donnée visualisée est associée à une couleur (rouge : spam ; bleu : non spam)
- Notes
  - 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)

Cahier des charges

L'application développée

- est basée de préférence sur Python/Django
- affiche les anomalies en 2D ou 3D, au choix
- inclut 1 interface graphique, utilisant de préférence la bibliothèque. D3JS : http://d3js.org/. L'interface graphique pourra être composée de 2 écrans :
  - Optionnel : chaque log est représenté, et est accessible directement par l'écran de détection d'anomalie
- inclue 1 bibliothèque de classification (SVM) ou de clustering (K-Means) des données
- 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 des algorithmes.

Dataset

Vous utiliserez le dataset de données de spam mis à disposition par les Hewlett-Packard Labs via le Machine Learning repository de l'UC Irvine :

- http://archive.ics.uci.edu/ml/datasets/Spambase