Université de Blida1

Département d'Informatique

Module : Intelligence artificielle L3 SIQ 2022/2023

Mini projet :

L'objectif de ce travail est de détecter les intrusions dans un réseau en analysant les paquets TCP/IP par des algorithmes d'apprentissage automatique. Les intrusions sont simulées dans un réseau LAN militaire réel. Ce dernier faisait l'objet de multiples attaques simulées. Une connexion est une séquence de paquets TCP commençant et se terminant à une certaine durée et dans laquelle les données circulent depuis une adresse IP source vers une adresse IP cible sous un protocole bien défini. En outre, chaque connexion est étiquetée comme normale ou anomalie (attaque/intrusion) avec exactement un type d'attaque spécifique. Dans ce travail, le type d'attaque est ignoré. Chaque enregistrement de connexion comprend environ 100 octets. Pour chaque connexion TCP / IP, 41 caractéristiques quantitatives et qualitatives sont obtenues à partir de données normales et d'attaques (3 caractéristiques qualitatives et 38 caractéristiques quantitatives). La variable 'class' caractérise le type de paquet et appartient à l'une des catégories suivantes :

- Normal
- Anomalie

L'acquisition de données s'est faite sur trois phases :

Phase 1:

Les données collectées au cours de la première phase sont stockées dans le fichier Phase1.csv

1.1. Faites une analyse exploratoire des données : (valeurs manquantes, données redondantes, transformations des données,..., etc.).

1.2. Appliquez un algorithme d'apprentissage supervisé pour détecter les anomalies.

1.3. Programmez une fonction qui permet de calculer le nombre de vrai positif, faux positif, vrai négatif et faux négatif dans un ensemble de données.

1.4. Calculez la précision, le rappel et le F1 score? en utilisant la validation croisée 5-folds et 30% des données pour le test.

1.5. Le modèle donne-il de bons résultats ? Les ingénieurs ont remarqué que les données contiennent des problèmes. Ils ont décidé alors de faire une deuxième acquisition de données. D'après votre analyse, de quel problème s'agit-il ?

Phase2:

Les données acquise dans la phase 2 sont stockées dans le fichier Phase2.csv

2.1. Faites une deuxième analyse exploratoire. Quels sont les changements apportés aux données ?

2.2. Est-ce que ce changement a un impact sur les performances du classificateur ? Si oui, expliquez le changement en justifiant votre réponse.

Phase3:

Afin d'augmenter le nombre des données de test et pour vérifier la stabilité des résultats obtenus, les ingénieurs ont procédé à une troisième acquisition de données. Toutefois, un problème technique est survenu et a empêché d'annoter les données avec le type de paquet (anomalie ou normal). Les données collectées durant cette phase sont sauvegardées dans le fichier Phase3.csv.

3. 1. Appliquez un algorithme d'apprentissage non supervisé sur le fichier Phase3.csv pour identifier les anomalies.

3.2. Mesurez les performances de cet algorithme.

3.3. Est-il possible de combiner vos travaux sur le fichier Phase3.csv et Phase2.csv afin d'obtenir un modèle robuste ? Si oui, proposez une approche et implémentez-la. Quelle sont les améliorations remarquées ?

Bon courage