Chapitre 4 : Principaux Datasets

Dr. ZAMOUCHE Djamila

Université A. MIRA - Bejaia

Faculté des Science Exactes

Département d'Informatique

Email : djamila.zamouche@univ-bejaia.dz

2024 / 2025

Table des matières

Objectifs

A l'issu de ce chapitre, l'étudiant sera capable de :

- Connaître les principaux ensembles de données (datasets) utilisés pour l'évaluation des IDS ;
- Différencier entre les principaux datasets de sécurité ;
- Choisir efficacement un dataset pour la mise en œuvre et l'évaluation d'un IDS.

Introduction

Bien qu'il n'existe pas de cadre de test standard capable d'évaluer de manière exhaustive les prototypes proposés, plusieurs ensembles de données (Datasets) ont été utilisés au cours des dernières années pour l'évaluation des IDS.

Principaux datasets de securité

1. Dataset DARPA 98

Le premier effort d'évaluation de la détection d'intrusion sponsorisé par le DARPA a été réalisé en 1998. L'évaluation comprenait deux parties : une évaluation hors ligne et une évaluation en temps réel. Les deux évaluations impliquent la mise en place de réseaux simulés pour générer du trafic réseau. Le réseau simulé comprend des segments intérieurs et extérieurs séparés par un routeur. Outre le trafic réseau réel, un trafic d'attaque est généré artificiellement. Les principaux services utilisés dans le réseau simulé sont HTTP, FTP, SMTP, POP, DNS, X, IRC, SWL/Telnet, SNMP, Time et Finger, qui couvrent la plupart des services fréquemment utilisés dans le réseau réel.

Le trafic contient 38 types d'attaques qui se répartissent en quatre catégories quatre catégories :

- Déni de service (DOS)
- De distant à utilisateur (R2L)
- Utilisateur vers root (U2R)
- Surveillances/Probing.

Sept semaines d'attaques basées sur le réseau au milieu de données normales sont collectées pour l'apprentissage. Les deux semaines de tests contiennent des attaques qui n'existent pas dans les données de l'apprentissage. Dans les expériences, deux types de données ont été collectées : 1) les données tcpdump capturées à partir du lien réseau et 2) les données d'audit du système.

2. Dataset DARPA 99

L'objectif principal de l'évaluation DARPA 99 était de mesurer la capacité du système de détection d'intrusion à détecter de nouvelles attaques, car il s'agissait d'un problème majeur découvert lors de l'évaluation 98. En conséquence, des stations de travail Windows NT ont été ajoutées au réseau simulé, et 17 nouvelles attaques ciblant les systèmes NT ont été insérées dans le trafic. L'autre changement majeur est l'ajout d'attaques internes.

Les données de reniflage interne, qui n'étaient pas utilisées dans l'évaluation de 98, sont fournies comme données de formation et de test afin de détecter ces attaques internes. En outre, l'évaluation des IDS est plus complète puisque la détection et l'identification des attaques sont évaluées.

3. Dataset KDD CUP 99

Depuis 1999, KDD CUP 99 est le jeu de données le plus utilisé pour l'évaluation des méthodes de détection des anomalies basées sur les réseaux. Cet ensemble de données est préparé par Stolfo et al. et est construit sur la base des données capturées dans le programme d'évaluation IDS de DARPA 98.

L'ensemble de données d'entraînement KDD se compose d'environ 4900000 vecteurs de connexion unique, chacun d'entre eux contenant 41 caractéristiques et étant étiqueté comme normal ou comme une attaque avec exactement un type d'attaque spécifique, tandis que l'ensemble de test contient environ 300000 échantillons avec un nombre total de 24 types d'attaques d'entraînement, plus 14 types supplémentaires dans l'ensemble de test uniquement. Le nom et la description détaillée des types d'attaques sont énumérés dans ∗

4. Dataset NSL-KDD

L'analyse de l'ensemble de données KDD montre qu'il y a deux problèmes importants dans l'ensemble de données qui affectent fortement la performance des systèmes évalués, et résulte en une très mauvaise évaluation des approches de détection d'anomalies. Pour résoudre ces problèmes, un nouvel ensemble de données est proposé, appelé NSL-KDD, qui consiste en des enregistrements sélectionnés de l'ensemble complet de données KDD. Cet ensemble de données est disponible publiquement pour les chercheurs et présente les avantages suivants par rapport à l'ensemble de données KDD original :

- Il n'inclut pas d'enregistrements redondants dans l'ensemble d'entraînement, de sorte que les classificateurs ne seront pas biaisés vers les enregistrements plus fréquents.
- Il n'y a pas d'enregistrements en double dans les ensembles de test proposés ; par conséquent, les performances des apprenants ne sont pas biaisées par les méthodes qui ont de meilleurs taux de détection sur les enregistrements fréquents.
- Le nombre d'enregistrements sélectionnés dans chaque groupe de niveau de difficulté est inversement proportionnel au pourcentage d'enregistrements dans l'ensemble de données KDD original. Par conséquent, les taux de classification des différentes méthodes d'apprentissage automatique varient dans une fourchette plus large, ce qui rend plus efficace l'évaluation précise des différentes techniques d'apprentissage.
- Le nombre d'enregistrements dans les ensembles de formation et de test est raisonnable, ce qui rend abordable l'exécution des expériences sur l'ensemble de l'ensemble sans avoir besoin de sélectionner aléatoirement une petite partie. Par conséquent, les résultats d'évaluation de différents travaux de recherche seront cohérents et comparables.

5. Dataset DARPA 2000

L'évaluation DARPA 2000 visait la détection d'attaques complexes qui contiennent plusieurs étapes. Deux scénarios d'attaque sont simulés dans l'évaluation 2000, à savoir LLDOS (Lincoln Laboratory Scenario (DDoS)) 1.0 et LLDOS 2.0. Les deux scénarios d'attaque sont exécutés sur plusieurs sessions de réseau et d'audit.

Ces sessions sont regroupées en quatre phases d'attaque : 1) sondage ; 2) effraction du système en exploitant une vulnérabilité ; 3) installation d'un logiciel DDOS sur le système compromis ; et 4) lancement d'une attaque DDOS contre une autre cible. LLDOS 2.0 est différent de LLDOS 1.0 dans le sens où les attaques sont plus furtives et donc plus difficiles à détecter, et donc plus difficiles à détecter. Comme cet ensemble de données contient un scénario d'attaque en plusieurs étapes, il est aussi couramment utilisé pour l'évaluation des méthodes de corrélation des alertes.

6. DEFCON 9

L'ensemble de données DEFCON 9 est un autre ensemble de données couramment utilisé pour l'évaluation des IDS. Ces données contiennent du trafic réseau capturé lors d'une compétition de hackers appelée Capture The Flag (CTF), dans laquelle les équipes concurrentes étaient divisées en deux groupes : les attaquants et les défenseurs. Le trafic produit pendant le CTF est très différent du trafic réseau du monde réel puisqu'il ne contient que du trafic intrusif sans aucun trafic de fond normal. En raison de cette lacune, cet ensemble de données est généralement utilisé pour l'évaluation des techniques de corrélation d'alertes. Ces ensembles de données sont des atouts précieux pour la communauté de la détection des intrusions. Cependant, ils souffrent tous du fait qu'ils ne sont pas de bons représentants du trafic du monde réel. L'ensemble de données du DARPA, par exemple, a été mis en doute quant au réalisme du trafic de fond car il est généré synthétiquement. Sur Outre la difficulté de simuler le trafic réseau réel, l'évaluation des IDS présente d'autres défis, notamment la difficulté de collecter les scripts d'attaque et les logiciels victimes, les exigences différentes pour tester les IDS basés sur les signatures ou sur les anomalies, les IDS basés sur le réseau ou sur l'hôte, etc.

Exercices

1. Exercice

[solution n°1 p. 8]

DEFCON 9 contiennent du trafic réseau capturé lors d'une compétition de hackers appelée Capture The Flag (CTF).

- ○ Vrai
- ○ Faux

2. Exercice

[solution n°2 p. 8]

Les avantages du dataset NSL KDD par rapport au dataset KDD original sont :

- ☐ Il n'inclut pas d'enregistrements redondants dans l'ensemble d'entraînement.
- ☐ Les données étant étiquetées comme normal ou comme une attaque avec exactement un type d'attaque spécifique.
- ☐ Il n'y a pas d'enregistrements en double dans les ensembles de test. par conséquent, les performances des apprenants ne sont pas biaisées par les méthodes qui ont de meilleurs taux de détection sur les enregistrements fréquents.
- ☐ Le nombre d'enregistrements dans les données d'entraînement et de test est raisonnable, ce qui rend abordable l'exécution des expériences sur l'ensemble sans avoir besoin de sélectionner aléatoirement une petite partie.

Solutions des exercices

Solution n°1

[exercice p. 7]

DEFCON 9 contiennent du trafic réseau capturé lors d'une compétition de hackers appelée Capture The Flag (CTF).

- (●) Vrai
- ( ) Faux

Solution n°2

[exercice p. 7]

Les avantages du dataset NSL KDD par rapport au dataset KDD original sont :

- [x] Il n'inclut pas d'enregistrements redondants dans l'ensemble d'entraînement.
- [ ] Les données étant étiquetées comme normal ou comme une attaque avec exactement un type d'attaque spécifique.
- [x] Il n'y a pas d'enregistrements en double dans les ensembles de test. par conséquent, les performances des apprenants ne sont pas biaisées par les méthodes qui ont de meilleurs taux de détection sur les enregistrements fréquents.
- [x] Le nombre d'enregistrements dans les données d'entraînement et de test est raisonnable, ce qui rend abordable l'exécution des expériences sur l'ensemble sans avoir besoin de sélectionner aléatoirement une petite partie.

Bibliographie

DARPA 1998 Dataset. (Disponible online :) https ://www.ll.mit.edu/r-d/datasets/1998- darpa-intrusion-detection-evaluation-dataset.

DARPA 1999 Dataset. (Disponible online :) https ://www.ll.mit.edu/r-d/datasets/1999- darpa-intrusion-detection-evaluation-dataset.

KDD CUP -1999 Dataset. (Disponible online :) https ://www.kaggle.com/datasets/galaxyh/kddcup- 1999-data.

NSL-kdd Dataset. (Disponible online :) https ://www.unb.ca/cic/datasets/nsl.html.

DARPA 2000 Dataset. (Disponible online :) https ://www.ll.mit.edu/r-d/datasets/2000- darpa-intrusion-detection-scenario-specific-datasets.