IDENTIFICATION D'INTRUSION SUR UN RÉSEAU INFORMATIQUE

Cette étude montre comment xtractis réussit à détecter une intrusion sur un réseau informatique militaire (US Air Force).

Dans le cas d'une alerte d'intrusion, un multiclassifieur flou xtractis permet de qualifier le type d'attaque pour accélérer la mise en place du traitement adéquat.

Résultats par :
xtractis® GENERATE 9.2.24581
upd 1807

TYPE DE MODÉLISATION

Modèle n°1 – Modèle de Classification : présence / absence d'une intrusion.

Modèle n°2 – Modèle de Multiclassification : qualification de l'intrusion (4 classes).

DONNÉES D'APPRENTISSAGE

source : Cyber Systems and Technology group of MIT Lincoln Laboratory, DARPA ITO, Air Force Research Laboratory [UCI Machine Learning Repository]

Dimension des données : 215 590 cas de référence (doublons retirés pour éviter l'overfitting).

Partitionnement : 71% pour l'apprentissage et 29% pour le test externe ETD.

118 prédicteurs potentiels décrivant la connexion au réseau (TCP dump data).

Les données d'apprentissage comportent 39.89% cas d'attaques.

Modèle n°1 – Sortie qualifiée avec 2 états : réseau sous le coup d'une intrusion (1) / pas d'intrusion (0)

Modèle n°2 – Sortie qualifiée avec 5 classes : Normal (60.11%) / DOS (35.71%) / PROBE (2.23%) / R2L5 (1.83%) / U2R (0.12%).

SOLUTION XTRACTIS

Modèle n°1

Grâce à leur intelligence collective et évolutive, les robots xtractis explorent 300 stratégies différentes d'apprentissage inductif, puis sélectionnent la stratégie la plus performante pour former un collège d'experts virtuels (EVI). Ce collège permettra de diagnostiquer la présence ou l'absence d'une attaque.

Dimension du collège : 20 modèles combinant 478 règles en tout (19 à 26 règles par modèle) et utilisant 117 variables prédictives (116 à 117 variables par modèle).

Le collège décide suivant un vote à l'unanimité des EVI, en tolérant le vote blanc.

Modèle n°2

Les robots xtractis explorent 1 500 stratégies différentes d'apprentissage inductif, puis sélectionnent la stratégie la plus performante pour former un collège d'EVI. Ce collège permettra de qualifier le type d'attaque.

Dimension du collège : 20 modèles combinant 709 règles en tout (32 à 36 règles par modèle) et utilisant 117 variables prédictives (115 à 117 variables par modèle).

L'agrégateur du collège est la moyenne des degrés de Possibilité normalisés.

CRITÈRES DE PERFORMANCE

La performance du collège de modèles est mesurée grâce à la validation croisée : 20 partitionnements aléatoires composés de 80% de points en training et 20% en validation. Les points de training sont utilisés pour créer le modèle, les points de validation sont utilisés pour évaluer la robustesse ou capacité prédictive du modèle et les points de l'ETD sont utilisés pour constater la performance réelle du modèle (prédiction sur des cas inconnus du modèle).

Les métriques de performance sont :

- la Sensitivité (le taux de vrais positifs)
- la Spécificité (le taux de vrais négatifs)
- le minimum entre Sensitivité et Spécificité (minSS) car la difficulté est détecter le maximum de situations d'attaque en minimisant les fausses alertes
- le PPV (Positive Predictive Value) : la chance qu'une prédiction de 1 faite par le collège corresponde à un vrai positif
- le NPV (Negative Predictive Value) : la chance qu'une prédiction de 0 faite par le collège corresponde à un vrai négatif
- une erreur globale de classification

RÉSULTATS

Les performances des modèles évaluées sur ces critères sont reportées dans les différentes matrices de confusion :

Modèle n°1

CB2 – intrusion (simple unanimity)	Classification error	Min. Sensitivity Specificity	Refused (xtractis utilise son droit de refus)
Accuracy /Training (300sr x 20sp x 80%)	0.11%	99.81%	1 420 (0.93%)
Robustness/Validation (300sr x 20sp x 20%)	0.23%	99.68%	791 (0.53%)
Real performance /External Testing	**0.13%**	99.71%	720 (1.14%)

Erreur de classification et MinSS

Actual class

Predicted class: intrusion	0	1		
0	90 860	112	99.88%	0.12%
1	50	60 002	0.08%	99.92%
	99.95%	0.19%		
	0.05%	99.81%		
Refused	722 (0.79%)	698 (1.15%)		

Accuracy/Training Confusion matrix

Actual class

Predicted class: intrusion	0	1		
0	89 992	192	99.79%	0.21%
1	157	59 528	0.26%	99.74%
	99.83%	0.32%		
	0.17%	99.68%		
Refused	416 (0.46%)	375 (0.62%)		

Robustness/Validation Confusion matrix

Actual class

Predicted class: intrusion	0	1		
0	43 274	56	99.79%	0.21%
1	27	19 069	0.26%	99.74%
	99.94%	0.29%		
	0.06%	**99.71%**		
Refused	403 (0.92%)	317 (1.63%)		

Real performance/External Testing Confusion matrix

Modèle n°2

CB2 – intrusion (mean normalized possibilities)	Classification error	Min. Sensitivity	Refused (xtractis utilise son droit de refus)
Accuracy /Training (1 500sr x 20sp x 80%)	2.91%	95.24%	0 (0.00%)
Robustness/Validation (1 500sr x 20sp x 20%)	3.16%	91.98%	0 (0.00%)
Real performance /External Testing	**2.76%**	91.03%	1 (0.00%)

Erreur de classification et MinSS

Actual class

Predicted class	Normal	Probe	Dos	U2R	R2L
Normal	42453	7	417	1	30
Probe	134	1310	14	0	2
Dos	64	2	16482	0	0
U2R	162	4	0	71	10
R2L	890	2	0	6	1084
Total	43703	1325	16913	78	1126
Unavailable	0	0	0	0	0

Real Performance / External Testing Confusion Matrix - Occurences

Actual class

Predicted class	Normal	Probe	Dos	U2R	R2L
Normal	**97.14%**	0.53%	2.47%	1.28%	2.66%
Probe	0.31%	**98.87%**	0.08%	0.00%	0.18%
Dos	0.15%	0.15%	**97.45%**	0.00%	0.00%
U2R	0.37%	0.30%	0.00%	**91.03%**	0.89%
R2L	2.04%	0.15%	0.00%	7.69%	**96.27%**

Real Performance / External Testing Confusion Matrix - Sensitivity Rates

Actual class

Predicted class	Normal	Probe	Dos	U2R	R2L
Normal	**98.94%**	0.02%	0.97%	0.00%	0.07%
Probe	9.18%	**89.73%**	**0.96%**	0.00%	0.14%
Dos	0.39%	0.01%	**99.60%**	0.00%	0.00%
U2R	65.59%	1.62%	0.00%	**28.74%**	4.05%
R2L	44.90%	0.10%	0.00%	0.30%	**54.69%**

Real Performance / External Testing Confusion Matrix - PPV Rates

COMPRENDRE LES RÉSULTATS

Modèle n°1

En situations réelles inconnues, xtractis fournit d'excellents résultats :

- les indicateurs de robustesse sont fiables puisque l'erreur réelle de classification (0.13%) est meilleure que celle estimée lors de la validation (0.23%) ;
- excellente Sensitivité : pour 10 000 cas d'intrusions avérées, xtractis en détecte 9 971;
- excellente Spécificité : pour 10 000 cas de non intrusion, xtractis en perçoit 9 994;
- Le minSS est très élevé (99.71%) confirmant la très bonne détection aussi bien des cas d'intrusion sur le réseau informatique que de cas de connexion autorisées. Selon le principe de précaution, les experts humains ont tendance à privilégier la Sensitivité au détriment de la Spécificité, d'où un grand nombre d'alertes inutiles ;
- une très faible erreur globale de classification : 0.13%. En matière de sécurité informatique, la moindre intrusion peut s'avérer fatale au système, il est donc nécessaire d'approcher le modèle parfait ;
- En situations opérationnelles, la confiance dans ce système décisionnel prédictif est validée par les très hautes valeurs du PPV et NPV : lorsque le système classe un nouveau dossier comme « intrusion », il a raison dans 99.86% des cas. Lorsqu'il classe un nouveau dossier comme « pas d'intrusion », il a raison dans 99.87% des cas.

Modèle n°2

Les indicateurs de robustesse sont fiables puisque l'erreur réelle globale de classification (2.76%) est inférieure à celle estimée lors de la validation (3.16%).

En situations réelles inconnues, xtractis fournit de très bons résultats dans la détection des connexions normales (Sensitivité=97.14%), à un niveau légèrement inférieur au modèle n°1 (99.94%) : ici on cherche surtout à caractériser le type d'intrusion détectée.

Le modèle s'avère être capable de très bien distinguer les attaques de type PROBE (Sensitivité=98.77%), DOS (Sensitivité=97.45%) et R2L (Sensitivité=96.27%), il reste bon pour détecter les attaques de type U2R (Sensitivité=91.03%)

L'erreur globale de classification semble assez faible mais reste 20 fois supérieure à celle du modèle n°1. Or, il est important de disposer de la plus faible erreur de classification, la moindre intrusion non détectée pouvant s'avérer fatale au système. Nous préconisons donc d'utiliser d'abord le modèle n°1 pour faire le tri entre attaque/pas d'attaque, puis le modèle de multiclassification pour caractériser l'attaque.

En situations opérationnelles, la confiance dans ce système décisionnel prédictif ne sera validée que pour les classes ayant des PPV importants : pas d'intrusion (98.94%), attaque de type DOS (99.60%) et attaque de type PROBE (89.73%). Cela veut dire que lorsque le système classe une nouvelle situation comme attaque de type DOS, il a raison dans 99.60% des cas, et ainsi de suite…

Toutefois, de nombreuses fausses alarmes sont déclenchées pour les prédictions d'attaque U2R et R2L : 65.59% de connexion sans intrusion sont perçus comme de potentielles attaques U2R et 44.90% de connexion sans intrusion sont perçus comme de potentielles attaques R2L. Du fait de la rareté de ces deux types d'attaque, ces fausses alarmes font ainsi tomber le PPV pour chacune de ces classes respectivement à 28.74% et 54.69%.

En réalité, la base de données de référence est déséquilibrée : il y a trop peu de cas d'attaque de type U2R par exemple pour pouvoir correctement les caractériser, d'autant que les différents types d'attaque regroupent un certains nombres de techniques d'attaques différentes (exemple pour le type d'attaque U2R : buffer_overflow, loadmodule, perl, …).

Pourquoi l'IA est-elle nécessaire en cybersécurité ?

À l'heure actuelle, la plupart des solutions de cybersécurité se fondent sur des règles expertes modélisant l'expertise d'ingénieurs qualifiés. C'est l'approche classique des systèmes experts des années 70 qui est donc employée pour la conception de tels systèmes.

Toutefois, les études de psychologie cognitive démontrent depuis les années 50 que le raisonnement humain conscient est limité à au plus 9 critères en simultané ; de ce fait, il ne pourra concevoir de règles décisionnelles mettant en jeu des dizaines voire des centaines de variables en interaction pour modéliser des comportements malveillants complexes (signaux faibles). En conséquence, de tels systèmes de cybersécurité seront dans le meilleur des cas sous-optimaux.

Afin de parer aux multiples cybermenaces orientées aussi bien vers le vol de données que le dysfonctionnement volontaire d'industries stratégiques (hôpitaux, réseaux d'énergie dont les centrales nucléaires, banques, assurances, avions, véhicules autonomes…), il devient nécessaire de faire appel à des IA capables de découvrir seules les comportements malveillants. À terme, ce sont des IA qui lanceront des attaques et assureront la défense cybersécuritaire. L'IA permettra de combler la pénurie mondiale d'ingénieurs sécurité, aussi bien en nombre qu'en niveau de qualification.

À partir d'un ensemble de logs, xtractis est capable de déterminer le comportement de l'attaquant (grâce au raisonnement inductif d'xtractis GENERATE) : ce modèle à base de règles floues explique de manière intelligible les relations causales correspondant à un certain type d'attaque ou à un comportement nominal. Les comportements seront généralement différents selon la nationalité de l'attaquant.

Ce modèle robuste d'attaquant virtuel est immédiatement exploité en défensif pour détecter en temps réel, si un log correspond à tel type d'attaque ou à un comportement nominal (et ceci grâce au raisonnement déductif d'xtractis PREDICT).

Lorsque l'attaquant a détecté qu'il a été détecté par le système défensif, il va changer son comportement pour tenter de nouvelles pénétrations. xtractis MONITOR pourra alors alerter sur l'augmentation de comportements anormaux et demander à GENERATE de relancer la découverte du nouveau comportement d'attaques, sans devoir passer plusieurs mois ou années de recherche.

L'attaquant aura toujours l'avantage s'il dispose de la même IA (généralement Open Source) que celle utilisée par le défenseur. C'est pour cette raison, qu'une IA efficiente et non publique telle xtractis procure un avantage important dans les domaines de la Défense, Sécurité et Cybersécurité.

Retrouvez plus d'études de cas sur www.xtractis.ai/casdusage.

