Rapport sur le stage de formation à l'étranger effectué du 21 Octobre 2018 au 20 Décembre 2018

Dans le laboratoire :

Lab-STICC UMR CNRS 6285, Université de Bretagne Occidentale, 29200 Brest

Stage Doctorat

Kherbache Meriem
Doctorante en deuxième année
Spécialité Informatique
Dirigée par : Dr AMROUN Kamal
Codirigé par : Dr ESPES David

Financé par l'université
Abderrahmane Mira Bejaia

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

Remerciements

Je tiens à commencer ce rapport de stage par des remerciements, à l'université de Bejaia de m'avoir permis d'effectuer ce stage, à ceux qui m'ont beaucoup appris au cours de ce stage, et même à ceux qui ont eu la gentillesse de faire de ce stage un moment très profitable.

Aussi, je remercie Monsieur David Espes, mon directeur de stage qui m'a formé et accompagné tout au long de ce stage avec beaucoup de patience et de pédagogie. Enfin, je remercie l'école doctorale de m'avoir permis d'effectuer un stage au sein de leur établissement au cours de ces deux mois.

1. Introduction

Meriem Kherbache 25 ans, doctorante à l'université Abderrahmane Mira en Algérie en cours d'inscription à la troisième année, mon travail de thèse concerne la cyber-sécurité des communications et plus précisément la réduction des faux positifs dans les mécanismes de détection d'intrusion comportementale. Ces travaux de thèse sont dirigés par Dr. Kamal AMROUN, maître de conférences (HDR) à l'université de Bejaia, et sont co-encadrés par Dr. David ESPES, maître de conférences à l'Université de Bretagne Occidentale.

Mes travaux de thèse concernent l'hybridation des méthodes de fouilles de données et d'apprentissage automatique pour la détection d'intrusions dans un réseau. Elle est dirigée par Dr. Kamal AMROUN – Maitre de Conférences (HDR) à l'université de Abderrahmane Mira.

Les IDSs peuvent avoir deux fonctionnements différents : analyse par signature et l'analyse comportementale. Pour pouvoir détecter les nouvelles attaques, les mécanismes de détection d'intrusion par signature ne suffisent plus. En effet, ces derniers permettent de détecter avec certitude les attaques déjà connues et répertoriées. Ces mécanismes de détection d'intrusion sont donc utiles pour limiter l'impact de mises à jour non effectuées. Cependant, ils ne permettent nullement de se protéger contre des techniques d'attaques qui sont soit non connues (nommées 0-day) ou soit très furtives et discrètes.

Pour pallier à ces défauts, l'analyse comportementale du trafic réseau et des usages est donc la principale méthode pour pouvoir détecter ces attaques avancées. Néanmoins, il ne faut pas sous-estimer la difficulté de construire un modèle exhaustif et représentatif de la réalité.

Notre objectif est de proposer des méthodes basées sur l'analyse comportementale en utilisant des algorithmes d'apprentissage automatique (machine learning) et de fouille de données (data mining) afin d'améliorer les taux de détection de différentes attaques et de réduire les taux de faux-positifs.

Ce rapport résume les travaux effectués durant ce stage. En premier lieu, nous présentons la problématique de notre thèse pour bonne compréhension des problèmes. En deuxième lieu, nous allons mentionner les différents travaux effectués durant ce stage qui se compose des différents objectifs, travaux et enfin les résultats obtenus. Enfin, on conclut notre rapport.

2. Problématiques de la thèse

Cette thèse vise à contribuer à l'amélioration des méthodes d'évaluation des systèmes de détection d'intrusion. Cette thématique se compose de trois problématiques :

- Réduction du nombre de caractéristiques d'un jeu de données : l'algorithme de Machine Learning utilise un ensemble de caractéristiques pour identifier si un trafic réseau est bénin ou malicieux. Les caractéristiques utilisées vont donc avoir un impact direct sur l'efficacité de l'algorithme à classifier correctement le trafic mais également

sur la rapidité de classification de ce dernier. En effet, il faut pouvoir trouver un sous-ensemble optimal de caractéristiques qui permettra de trouver le meilleur compromis entre efficacité et rapidité.

- Identification d'attaques complexes à signaux faibles : avec l'amélioration des mécanismes de sécurité actuellement déployés, les attaques actuelles essaient d'être le plus discrètes possibles. Pour cela, ces attaques vont être réalisées sur une longue période temporelle et ressembler le plus possible à un trafic bénin. Actuellement les algorithmes de Machine Learning détectent particulièrement bien les attaques à forts signaux (tels que les attaques par Dénis de Service ou Dénis de Service Distribués). Il est donc nécessaire de faire évoluer les algorithmes de classification afin qu'ils puissent détecter de manière efficace les attaques complexes.
- Utilisation d'algorithmes d'apprentissage non-supervisés : il existe deux classes d'algorithmes d'apprentissage : les algorithmes supervisés et les algorithmes non-supervisés. Durant la phase d'entrainement, les algorithmes supervisés reposent sur un étiquetage précis du trafic (bénin, malicieux) pour construire leur modèle de détection. Avoir accès à un trafic étiqueté est souvent une tâche fastidieuse et onéreuse car elle repose sur la connaissance d'experts en sécurité. De même, la diversité des environnements de production rend peu générique de telles approches. A contrario, les approches non-supervisées reposent sur la similarité du trafic pour les catégoriser sans avoir recours à des informations complémentaires lors de l'entrainement du modèle. Une telle souplesse permet de ne pas avoir recours à des experts onéreux et de pouvoir s'adapter facilement à des environnements de production très différents. Cependant, l'efficacité limitée de ces algorithmes font qu'ils sont utilisés uniquement dans des environnements très spécifiques. Afin de disséminer au mieux la détection d'intrusion par analyse comportementale et d'en maîtriser les coûts d'exploitation, il est essentiel de pouvoir reposer sur des algorithmes non-supervisés.

3. Travaux effectués durant le stage

3.1. Objectifs du stage

Durant ce stage, mes travaux se focalisent préalablement sur les deux premiers verrous mentionnés précédemment et proposé une démarche rigoureuse couvrant l'ensemble des étapes de l'évaluation d'un IDS.

Durant la fin de ma deuxième année de thèse, j'ai proposé un algorithme de réduction de caractéristiques de type Wrapper. Cet algorithme est donc composé de deux phases : identification des caractéristiques pour chaque type d'attaque et sélection du sous-ensemble quasi-optimal de caractéristique.

Notre méthode a été appliquée sur un jeu de données très connu surnommé NSL-KDD, dans la littérature des IDSs. Les résultats obtenus montrent que la méthode de sélection de caractéristiques est robuste et permet une amélioration conséquente de l'efficacité des

algorithmes de classification. En effet, il nous a été permis de détecter avec un très faible taux de faux-positifs, des attaques particulièrement discrètes.

Afin d'enrichir et de s'assurer de l'efficacité de la méthode, nous avons réalisé des tests complémentaires sur différents jeu de données différents. Pour cela nous avons fixé les objectifs suivants :

1. Tester la méthode proposée sur différents jeux de données (ISCX2012, CICIDIS2017), pour montrer la faisabilité de notre approche car ces données sont de tailles importantes (CICIDS2017 **Trafic normal** : 180000000 trames). Cela nécessite une puissance matérielle puissante,
2. Adaptation des algorithmes de deuxième année de thèse en fonction des résultats obtenue dans 1.
3. Rédaction d'articles.
4. Proposition d'une nouvelle méthode de sélection de caractéristiques basée sur la variance et l'algorithme hiérarchique.

3.2. Travaux effectués

Afin d'atteindre les objectifs définis préalablement, durant ce stage, j'ai pu effectuer trois tâches. Dans un premier temps, nous avons commencé par traiter le jeu de données le plus récent CICIDS 2017, qui est un jeu de données vraiment complexe. Il se compose de 80 caractéristiques avec 12 différentes attaques (DDOS, Heartbleed, PortScan...etc.) d'une taille qui varie entre 36 à 80000 trames par attaque et d'un trafic normal de 1800000 trames. Ce dernier nécessite une capacité mémoire énorme et une machine puissante afin d'adapter la méthode proposée pour le calcul de la variance et de la matrice de covariance.

Par la suite, le besoin de faire l'apprentissage avec différents sous-ensembles de caractéristiques consomme énormément de temps. Pour cela j'ai eu accès à une machine de 256Go de capacité mémoire afin d'appliquer la méthode proposée sur ce jeu de données. La machine m'a permis d'importer différentes attaques et d'appliquer ma méthode sur ces-dernières.

Dans un deuxième temps, mes papiers de recherche sont actuellement dans une phase de finalisation. On a pu avancer dans la rédaction des articles. Plusieurs séances de travail ont été effectuées avec mon co-directeur de thèse monsieur ESPES David afin de discuter sur le plan de l'article et sur la structuration du papier.

Finalement, ce stage nous a permis de discuter sur plusieurs propositions dans le contexte de la sélection de caractéristiques. Suite aux résultats obtenus avec le nouveau jeu de données, la méthode développée durant ma deuxième année de thèse doit être adaptée. En effet, cette méthode requérait de fixer un seuil pour choisir les caractéristiques dont la variance est au-delà. On avait fixé ce seuil à 95% pour NSL-KDD. Cependant ce seuil ne fonctionne strictement pas avec CICIDS-2017.

Nous avons donc proposé une nouvelle méthode de sélection de caractéristiques basée sur la variance des caractéristiques et l'algorithme hiérarchique. Cette méthode permet donc de déterminer automatiquement ce seuil. Cette adaptation permet d'améliorer l'efficacité de la méthode en la rendant totalement automatique et indépendante de toute intervention d'un expert scientifique. Ces améliorations ont été appliquées sur le jeu de donnée NSL-KDD et CICIDS-2017. Les premiers résultats obtenus sont encourageants et permettent de valider son fonctionnement.

En résumé, ce stage m'a permis de traiter les différents jeux de données vu la disponibilité des machines puissantes. L'accès à un cluster dédié a permis de faire les différents apprentissages. J'ai également finalisé la rédaction de l'article avec mon co-directeur de thèse. Mes travaux de recherches vont prochainement être soumis dans un article de journal et de conférence internationale.

4. Conclusion

J'ai effectué un stage d'une durée de deux mois du 21 Octobre 2018 au 19 décembre 2018. Le stage est effectué au sein du Lab-STICC UMR CNRS 6285, Université de Bretagne Occidentale.

Ce stage m'a permis d'avancer sur deux plans, sur l'état d'avancement de ma thèse, pour cela j'ai pu rejoindre le laboratoire LabSTICC afin de travailler avec mon co-directeur de thèse Monsieur David ESPES, le laboratoire m'a donné accès à une machine de forte capacité mémoire afin de traiter mes différents jeu de données, j'ai pu également finaliser la rédaction de mon article et enfin une méthode de sélection de caractéristiques a était proposée et prouver son efficacité sur le jeu de données NSL-KDD.

Sur le plan personnel, en effet j'ai pu acquérir une certaine autonomie, et un goût prononcé pour le travail. Cela m'a aussi permis de développer une certaine aisance quant à mener à bien un projet, me voir confier des responsabilités, ainsi que d'analyser un problème donné et chercher des solutions adéquates. Finalement, mon séjour m'a permis de rencontrer différents chercheurs de différentes nationalités (doctorants, maître de conférences et professeurs) œuvrant dans des domaines proches de mes intérêts de recherche et sujet de thèse. Ces rencontres m'ont évidemment été très utiles sur le plan académique.

Visa du laboratoire d'accueil

Le Directeur du Lab STICC-UBO

Emanuel RADOI

Signature

Kherbache Meriem

Université de Bretagne Occidentale
Lab-STICC UMR CNRS 6285
6 Av. le Gorgeu - CS 93837
29238 BREST CEDEX 3 - France
Tél. 33(0)2 98 01 61 26