DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Ons AOUEDI

Machine Learning-Enabled Network Traffic Analysis

Thèse présentée et soutenue à Nantes, le 02 Décembre 2022
Unité de recherche : Laboratoire des Sciences de Numérique de Nantes (LS2N), UMR 6004

Rapporteurs avant soutenance :

Adlen KSENTINI	Professeur, Eurecom, Sophia Antipolis, France
Sonia BEN MOKHTAR	Directrice de Recherche CNRS/INSA LYON, France

Composition du Jury :

Président :	Yassine HADJADJ-AOUL	Professeur, Université de Rennes I, France
Examinateurs:	Adlen KSENTINI	Professeur, Eurecom, Sophia Antipolis, France
	Sonia BEN MOKHTAR	Directrice de Recherche CNRS/INSA LYON, France
	Yusheng JI	Professeure, National Institute of Informatics, Japon
Dir. de thèse :	Benoît PARREIN	Maître de Conférences HDR, Nantes Université
Co-encadrant. de thèse :	Kandaraj PIAMRAT	Maître de Conférences, Nantes Université

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

Titre : Analyse du trafic réseau basée sur l'apprentissage automatique

Mot clés : Apprentissage automatique, Apprentissage fédéré, analyse du trafic

Résumé : L'Internet des Objets entraînent par son nombre de terminaux une explosion du trafic de données. Pour augmenter la qualité globale de réseau, il est possible d'analyser intelligemment le trafic réseau afin de détecter d'éventuel comportement suspect ou malveillant. Les modèles d'apprentissage automatique et d'apprentissage profond permettent de traiter ce très grand volume de données. Néanmoins, il existe certaines limites dans la littérature, notamment la confidentialité des données, le surapprentissage (manques de diversité dans les données) ou tout simplement le manque de jeu de données labélisées. Dans cette thèse, nous proposons de nouveaux modèles s'appuyant sur l'apprentissage automatique et l'apprentissage profond afin de traiter une grande quantité de données tout en préservant la confidentialité. Notre première approche utilise un modèle d'ensemble. Les résultats montrent une diminution du surapprentissage, tout en augmentant de 10% la précision comparé à des modèles de l'état de l'art. Notre seconde contribution s'attache aux problèmes de disponibilité des données labélisées. Nous proposons un modèle d'apprentissage semi-supervisé capable d'améliorer la précision de 11% par rapport à un modèle supervisé équivalent. Enfin, nous proposons un système de détection d'attaque s'appuyant sur l'apprentissage fédéré. Nommé FLUIDS, il permet de réduire la surcharge réseau de 75% tout en préservant de très haute performance et la confidentialité.

Title: Machine Learning-Enabled Network Traffic Analysis

Keywords: Machine Learning, Federated Learning, traffic analysis

Abstract: Recent development in network communication along with the drastic increase in the number of smart devices leads to an explosion in data generation. To this end, intelligent network traffic analysis can help to understand the behavior of connected smart devices and applications as well as provides defense against cyber-attacks. In this line, Machine Learning (ML) and Deep Learning (DL) models have the ability to model and uncover hidden patterns using training data or environment. Despite their benefits, major challenges need to be addressed such as model generalization (due to model overfitting), lack of label (due to the difficulty to label all the data), and privacy (due to recent regulations). In this thesis, new ML/DL-based models are proposed for tackling these challenges. The first contribution focuses on improving the generalization and classification performance by proposing an ensemble blending model. The simulation results show that the accuracy of the proposed ensemble model is 10%, better than some state-of-the-art models. Second, a semi-supervised model has been proposed and the experiment results show that unlabeled data boost the classification accuracy by 11% in comparison to its supervised version. Finally, a Federated Learning (FL) based Intrusion Detection System (IDS) has been proposed. It allowed the clients to learn an efficient intrusion detection model without the need to label their local data as well as to achieve high classification performance and improvement in terms of communication overhead (reduction by almost 75%).