



Cải thiện hiệu suất phát hiện tấn công thông qua lựa chọn đặc trưng và so sánh mô hình trên dữ liệu RT-IoT2022

Sinh viên thực hiện: Tạ Hồng Quý GVHD: TS. Đỗ Như Tài

INTRODUCTION

Problems definition:

- Input:** Tabular data (RT-IoT2022) gồm 12 lớp phân loại
- Output:** Nhãn lớp dự đoán tương ứng với mỗi mẫu dữ liệu

Challenge:

- Dữ liệu mất cân bằng nghiêm trọng
- Số lượng đặc trưng lớn

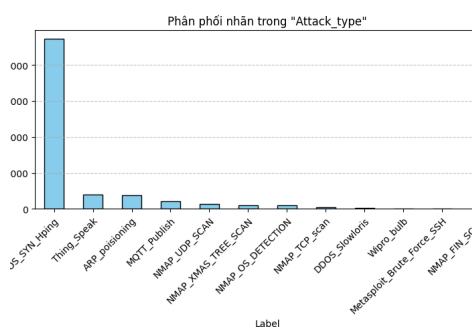
Mục tiêu:

- Tìm ra đặc trưng quan trọng giúp mô hình tăng hiệu suất

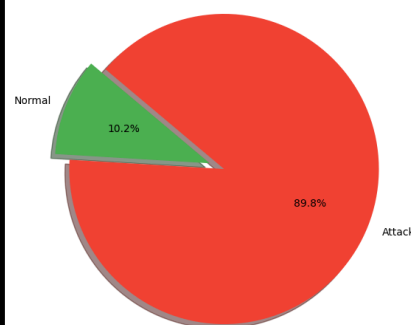
DATASET

- Quantity: Thu thập từ các thiết bị IoT. Chia dữ liệu thành hai phần train(80%) và test(20%).
- Data source: RT-IoT2022

[RT-IoT2022 - Kho lưu trữ máy học UCI --- RT-IoT2022 - UCI Machine Learning Repository](#)

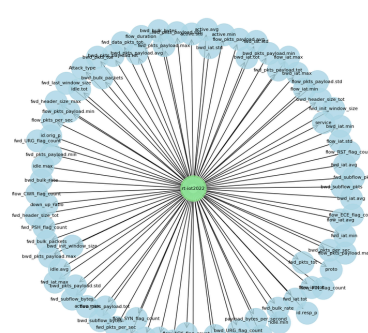


Phân bố dữ liệu: Bình thường vs Tấn công của dữ liệu IoT



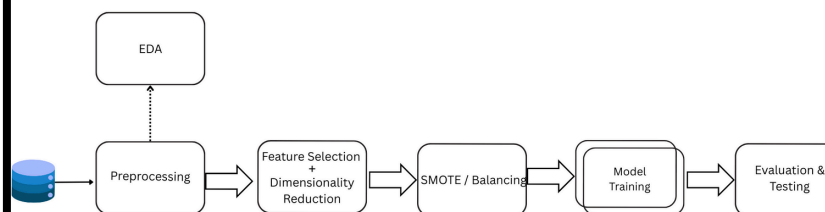
Nhãn attack gấp 9 lần normal

Feature of RT-IoT2022 Dataset



Số lượng đặc trưng lớn(85)

PROPOSED METHOD



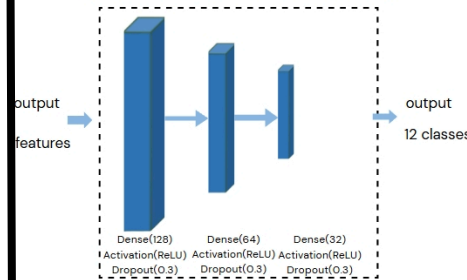
Model training

Mô hình học máy truyền thống

- 1.KNN
- 2.LinearSVC
- 3.XGBoost
- 4.Logistic Regression
- 5.Random Forest

Mô hình neural network MLP

Multilayer Perceptron Architecture



SMOTE/BALANCE

Attack Type	Original Count	Attack Type	Updated Count
DOS_SYN_Hping	75762	DOS_SYN_Hping	75762
Thing_Speak	6483	Thing_Speak	6483
ARP_poisoning	6172	ARP_poisoning	6172
MQTT_Publish	3275	MQTT_Publish	3275
NMAP_UDP_SCAN	2101	NMAP_UDP_SCAN	2101
NMAP_XMAS_TREE_SCAN	1626	NMAP_XMAS_TREE_SCAN	1626
NMAP_OS_DETECTION	1607	NMAP_OS_DETECTION	1607
NMAP_TCP_scan	782	NMAP_TCP_scan	782
DDOS_Slowloris	434	DDOS_Slowloris	434
Wipro_bulb	195	DDOS_Slowloris	500
Metasploit_Brute_Force_SSII	31	Metasploit_Brute_Force_SSII	500
NMAP_FIN_SCAN	25	NMAP_FIN_SCAN	500

Original trainset

SMOTE for trainset

EXPERIMENTS

phase 1: Dữ liệu chưa giảm chiều (Baseline)

phase 2: Dữ liệu đã giảm chiều (feature selection + ngưỡng tương quan)

phase 3: Tối ưu bằng RandomizedSearchCV/GridSearchCV

RESULTS

Bảng thực nghiệm trên của các giai đoạn

Mô hình	F1-score	Mô hình	F1-score
LinearSVC	0.840	LinearSVC	0.724
XGBoost	0.954	XGBoost	0.964
Logistic Regression	0.825	Logistic Regression	0.748
KNN	0.912	KNN	0.929
Random Forest	0.961	Random Forest	0.960
MLP	0.808	MLP	0.790

dữ liệu chưa giảm chiều

dữ liệu đã giảm chiều

Mô hình	F1-score	Mô hình	F1-score
XGBoost	0.950	XGBoost	0.956
KNN	0.920	KNN	0.920
Random Forest	0.951	Random Forest	0.962

RandomizedSearchCV

GridSearchCV