An effective genetic algorithm-based feature selection method for intrusion detection systems

Nhóm 14

Giảng viên: Đỗ Hoàng Hiển

Lớp: NT204.O21.ANTT

Danh sách thành viên

21522800

Nguyễn Long Vũ

21520911

Bùi Quốc Huy

21522735

Bùi Đức Anh Tú

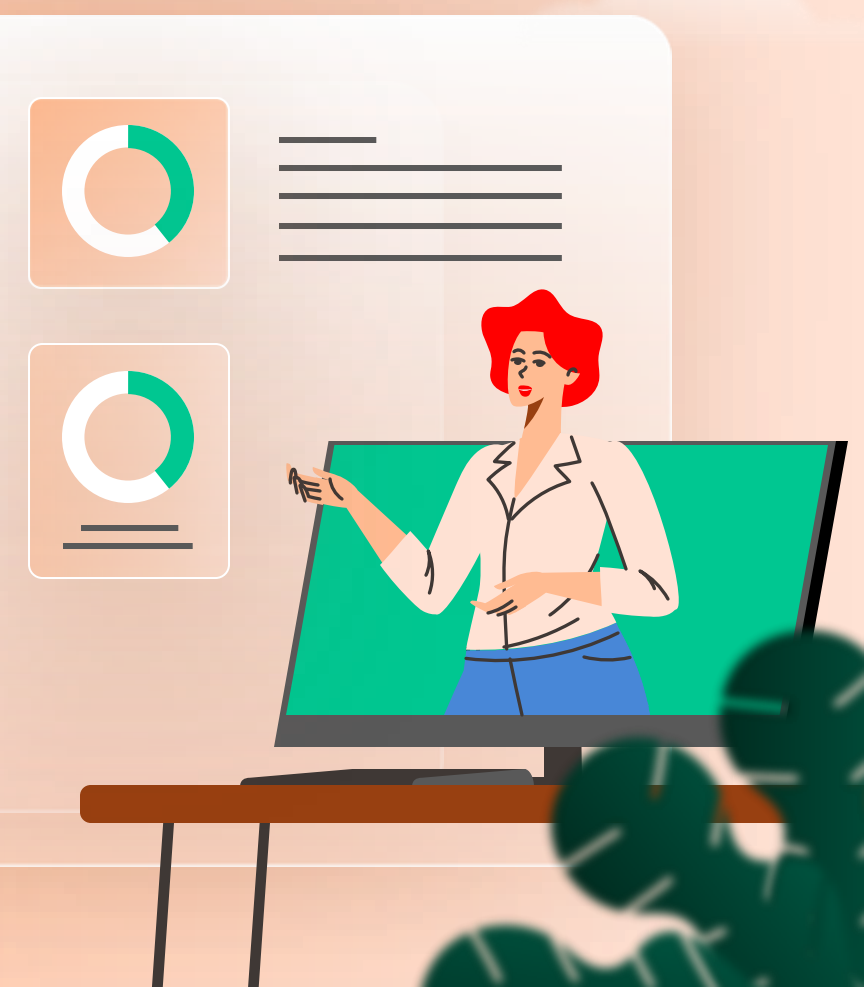
21522067

Lê Huy Hiệp

Nội dung trình bày

.01

Tổng quan đề tài



Tổng quan đề tài

Học máy đang được sử dụng rộng rãi trong việc phát triển các hệ thống phát hiện xâm nhập (IDS) để phát hiện các loại tấn công đã biết và biến thể của chúng, cũng như các loại tấn công chưa biết thông qua việc phân tích lưu lượng mạng.

Tuy nhiên sự tăng “chiều” (đặc trưng) của dữ liệu đã gây ra nhiều ảnh hưởng tiêu cực đối với hiệu suất của các thuật toán máy học. Nguyên nhân chính là do các thuộc tính của gói tin ngày càng tăng nhưng không phải thuộc tính nào cũng có ý nghĩa trong việc phân loại gói tin.

Tổng quan đề tài

Giảm “chiều” dữ liệu bằng cách loại bỏ các thuộc tính không liên quan và trùng lặp.

Giải pháp được đề xuất là “lựa chọn đặc trưng dựa trên thuật toán di truyền” để tìm ra các đặc trưng tối ưu của dataset, với mong muốn là khi huấn luyện với các đặc trưng tối ưu trong dataset đó sẽ đưa ra được các classifier có hiệu suất tốt hơn.

.02

Giải pháp được đề xuất

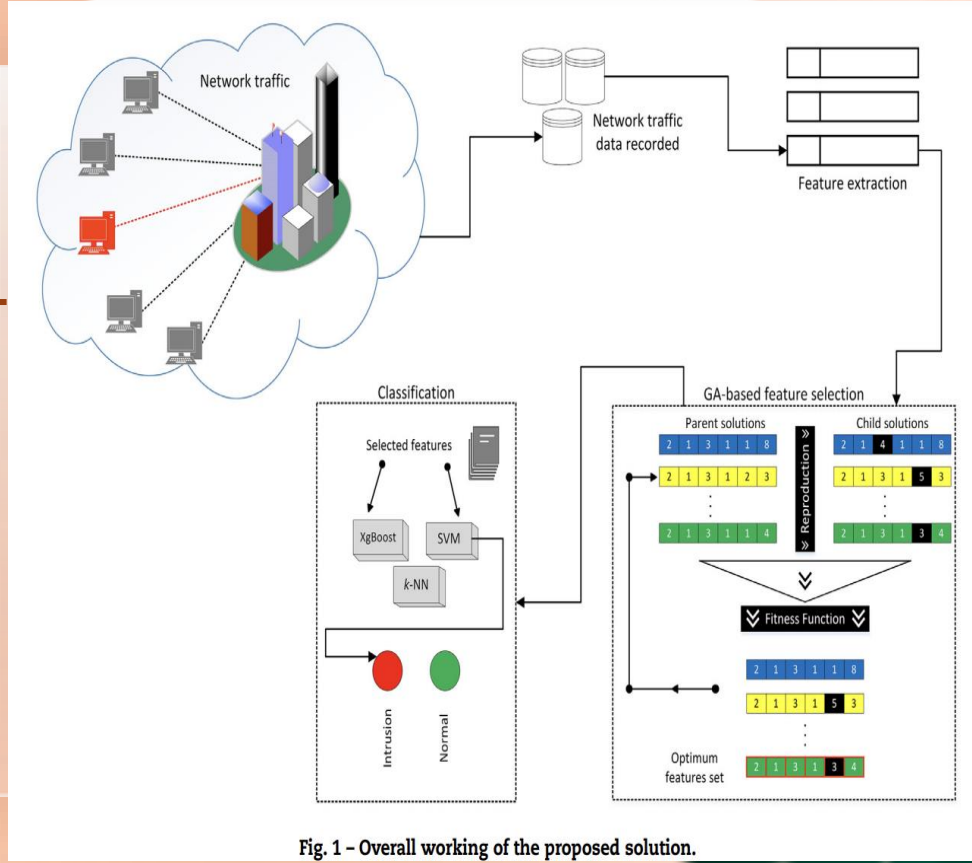
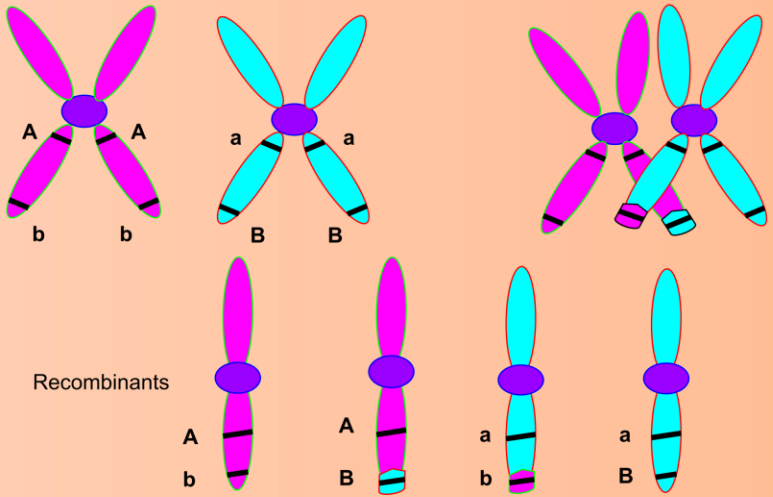


Fig. 1 – Overall working of the proposed solution.

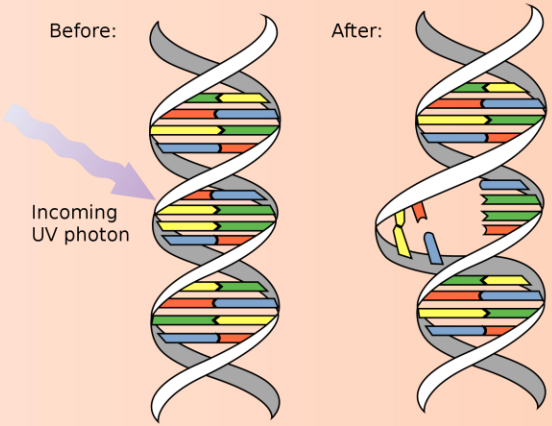
Phương pháp lựa chọn đặc trưng dựa trên GA (thuật toán di truyền)

- Khái niệm thuật toán di truyền: Là thuật toán dựa trên hiện tượng “di truyền” trong sinh học. Cụ thể, mô phỏng 2 quá trình:

+ Lai chéo (Trao đổi chéo)



+ Đột biến



Phương pháp lựa chọn đặc trưng dựa trên GA (thuật toán di truyền)

- Phương pháp lựa chọn đặc trưng dựa trên GA:

+ Tạo quần thể ban đầu

+ Tạo thế hệ:

* Lai chéo
* Đột biến
* Đưa ra "nhiễm sắc thể" tối ưu thông qua giá trị fitness

+ Tạo thế hệ mới

Phương pháp lựa chọn đặc trưng dựa trên GA (thuật toán di truyền)

- Phương pháp lựa chọn đặc trưng dựa trên GA:

+ *Tạo quần thể ban đầu:* Tạo ra quần thể với các cá thể là các "nhiễm sắc thể" (mảng chứa các đặc trưng ngẫu nhiên của dataset).

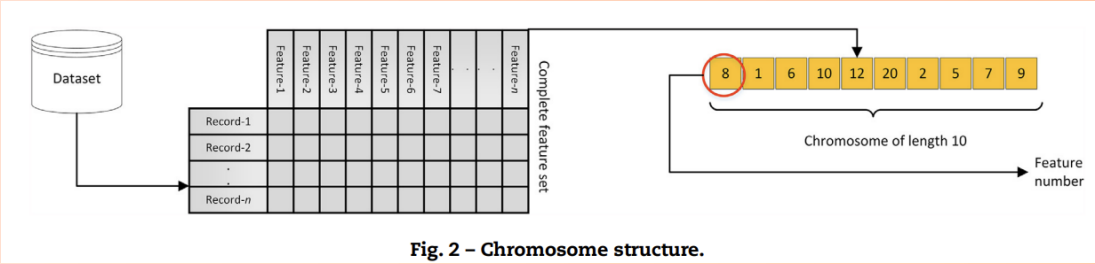


Fig. 2 – Chromosome structure.

Phương pháp lựa chọn đặc trưng dựa trên GA (thuật toán di truyền)

- Phương pháp lựa chọn đặc trưng dựa trên GA:

+ *Tạo thế hệ:*

* *Lai chéo:* "Cắt" nhiễm sắc thể 1 theo tỷ lệ cho trước và nối nó với phần còn lại của nhiễm sắc thể 2.

* *Đột biến:* Xét từng gen trên nhiễm sắc thể và dựa trên tỷ lệ cho trước để biến đổi nó 1 cách ngẫu nhiên.

* *Đưa ra "nhiễm sắc thể" tối ưu thông qua giá trị fitness:* Dùng hàm fitness để tính giá trị fitness cho các nhiễm sắc thể mỗi khi có quần thể mới được tạo ra. Nhiễm sắc thể có giá trị fitness cao nhất

+ *Tạo thế hệ mới:* Tạo tới khi "nhiễm sắc thể" tối ưu của thế hệ mới có giá trị fitness không cao hơn "nhiễm sắc thể" tối ưu của thế hệ cũ.

Phương pháp lựa chọn đặc trưng dựa trên GA (thuật toán di truyền)

- Phương pháp lựa chọn đặc trưng dựa trên GA:

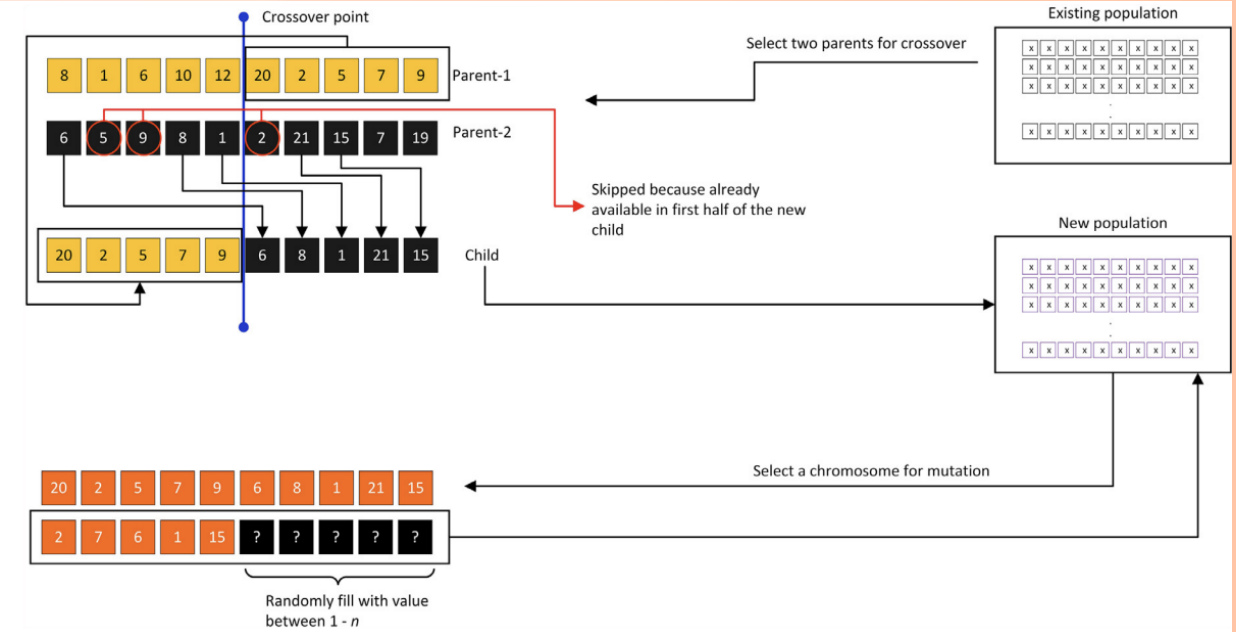


Fig. 3 – Crossover and mutation operations.

Phương pháp lựa chọn đặc trưng dựa trên GA (thuật toán di truyền)

- Hàm fitness: là hàm đánh giá "độ phù hợp" của NST (nhiễm sắc thể) thông qua 1 công thức do nhóm tác giả đề ra. Với $Corr_{avg}$ là trung bình tương quan của các đặc trưng được chọn, $Corr^{t}_{avg}$ là trung bình biến đổi tương quan, F_i là giá trị fitness của NST thứ i, A_i là độ chính xác có từ việc huấn luyện mô hình cho trước với các đặc trưng trong NST thứ i, M_i là ma trận tương quan của NST thứ i.

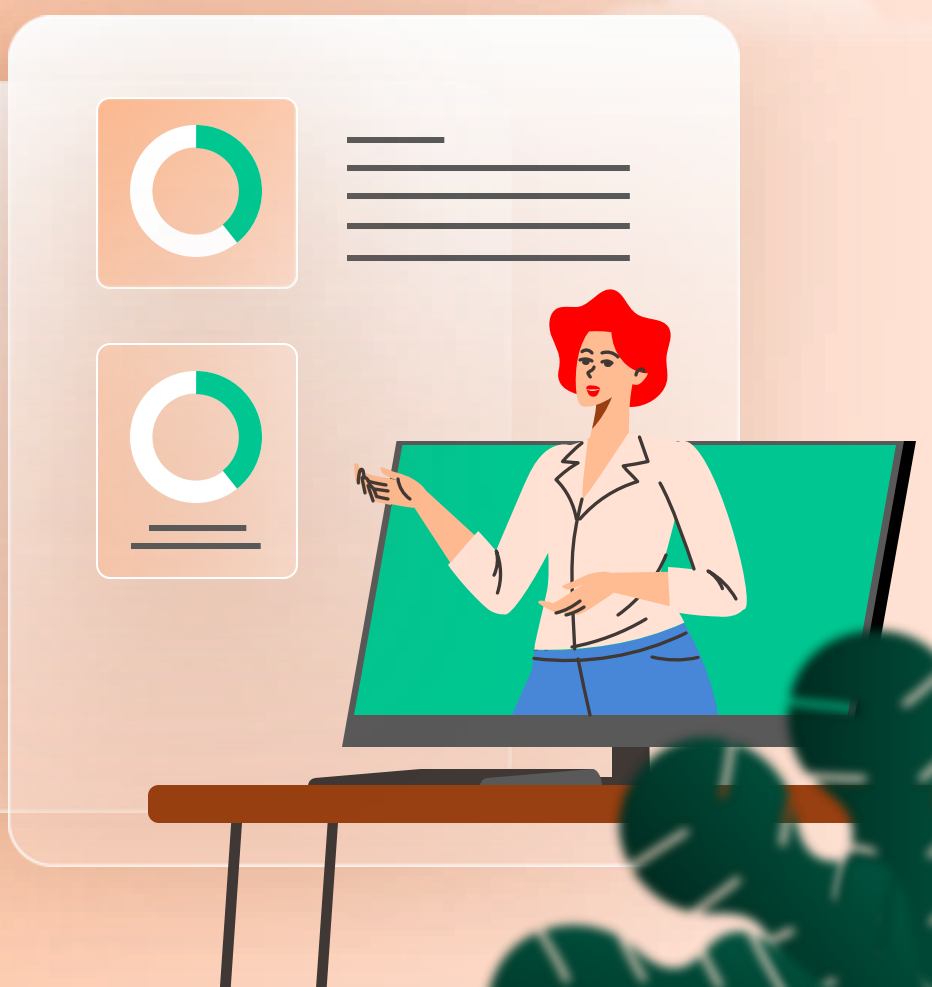
$$Corr_{avg} = \frac{Sum\ (s)\ of\ values\ above\ the\ diagonal}{Number\ of\ Values} \quad (3)$$

$$Corr^{t}_{avg} = (1 - Corr_{avg}) \quad (4)$$

$$F_i = \frac{A_i + (1 - M_i)}{2} \quad (5)$$

.03

Thiết kế giải pháp



- Chuẩn bị tập dữ liệu: CIRA-CIC-DOHBrw-2020, UNSW-NB15 và Bot-IoT 5%

+ Nạp bộ dữ liệu

+ Mã hóa theo nhãn sử dụng LabelEncoder

+ Sử dụng SimpleImputer để xử lý các giá trị rỗng

+ Sử dụng MinMaxScaler để Standardlize và Scaling

+ Phân chia train – test theo tỷ lệ 5 – 5

+ Cân bằng train set một cách thủ công với mục tiêu là số thực thể các lớp bằng nhau với kích thước train set sau cân bằng bằng với kích thước dataset.

- Lựa chọn đặc trưng dựa trên GA (GbFS):

+ Tạo quần thể với 100 NST, độ dài NST là 10

+ ***Hàm fitness***:

* *Ma trận tương quan:* sử dụng np.corrcoef(rowvar=False)
* *Trung bình tương quan:* (tổng ma trận – tổng đường chéo) / (kích thuớc ma trận – độ dài đường chéo)
* *Độ chính xác:* độ chính của mô hình cho trước khi huấn luyện với NST
* *Giá trị fitness:* (độ chính xác + (1 – trung bình tương quan của NST)) / 2

+ ***Chọn "cha mẹ" hay NST 1, 2 theo chiến thuật roulette wheel selection***: tạo mảng các NST được sắp xếp theo chiến thuật roulette wheel

+ ***Lai chéo***: tỷ lệ lai chéo = 0.5, tức cắt nửa sau của NST 1 gắn với nửa đầu của NST 2

+ ***Đột biến***: tỷ lệ đột biến = 0.5, tức có 50% tỷ lệ phần tử của NST được đổi ngẫu nhiên

+ ***Tạo thế hệ***: tạo quần thể ban đầu >> tính các giá trị fitness >> lưu NST tối ưu của thế hệ (fitness lớn nhất) >> vòng lặp lai chéo, đột biến >> trả về NST tối ưu nhất

* *Vòng lặp lai chéo, đột biến:* Chọn "cha mẹ" >> lai chéo >> tính fitness >> lưu NST tối ưu của thế hệ >> chọn "cha mẹ" >> đột biến >> tính fitness >> lưu NST tối ưu của thế hệ (nếu có) >> nếu NST tối ưu của thế hệ >> NST tối ưu thì lặp tiếp, nếu không thì dừng

- Tiến hành huấn luyện bộ phân loại với 10 đặc trưng tối ưu tìm được và với tất cả đặc trưng: SVM, k-NN, XgBoost

- Ghi nhận kết quả kiểm tra từ bộ phân loại và so sánh đánh giá.

Then

Select parents

Crossover

Fitness compute

Save gen's best chromo

Select parents

Mutation

Fitness compute

Initialize population

Save gen's best chromo

Continue if has new best chromo

Else

Best features

Dataset

Features

Model

Classifier

$$Corr_{avg} = \frac{Sum\ (s)\ of\ values\ above\ the\ diagonal}{Number\ of\ Values} \quad (3)$$

$$Corr^{t}_{avg} = (1 - Corr_{avg}) \quad (4)$$

$$F_i = \frac{A_i + (1 - M_i)}{2} \quad (5)$$

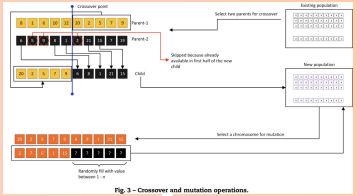


Fig. 3 – Crossover and mutation operations.

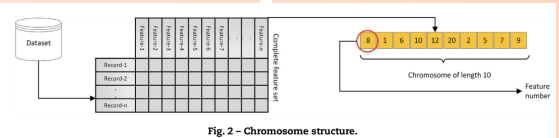
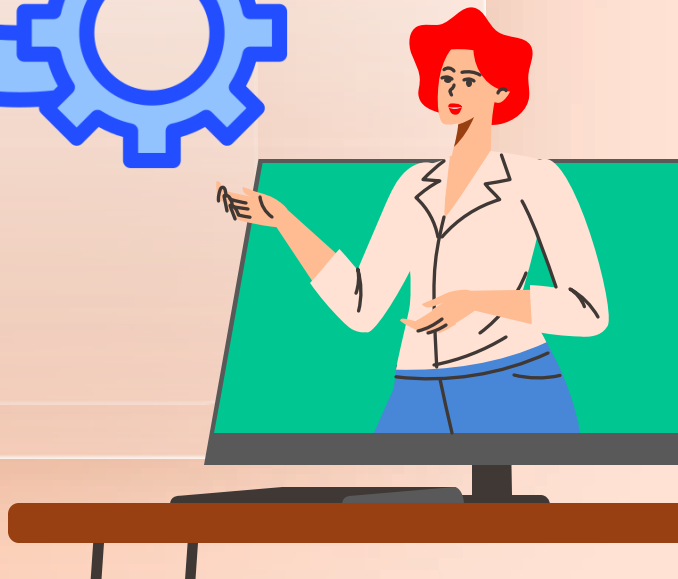


Fig. 2 – Chromosome structure.

.04

Datasets



CIRA-CIC-DOHBrw-2020

- Thống kê về dataset của bài báo:

Table 2 – Datasets summary.

	CIRA-CIC-DoHBrw-2020	Bot-IoT	UNSW NB-15
No. of features	34	29	49
No. of classes	4	5	10
No. of Samples	~1.4 million	~3 million	~0.25 million
X			

Table 4 – Attack classes in CIRA-CIC-DoHBrw-2020.

Classification of attack	No. of records	Attack name
DoH	269643	DNS over HTTPs
None-DoH	897493	None DNS over HTTPs
Benign-DoH	19807	Benign DNS over HTTPs
Malicious	249836	Malicious

- Thống kê về dataset của nhóm:

```
[3]  # Number of features
     print('Number of features: ', X.shape[1])

     Number of features:  34
```

```
# Record per class
print('Record per class:\n',data.groupby('Label').size())
print('\nSum:\t\t',data['Label'].size)

Record per class:
 Label
Benign        19807
DoH          269643
Malicious    249836
NonDoH       897493
dtype: int64

Sum:             1436779
```

UNSW-NB15

- Thống kê về dataset của bài báo:

Table 2 – Datasets summary.

	CIRA-CIC-DoHBrw-2020	Bot-IoT	UNSW NB-15
No. of features	34	29	49
No. of classes	4	5	10
No. of Samples	~1.4 million	~3 million	~0.25 million
X			

Table 8 – Per class records in UNSW NB-15 dataset.

Classification of attack	No of records
Analysis	677
Backdoor	577
DoS	4089
Exploits	7061
Fuzzers	12,062
Generic	5016
Normal	31,395
Reconnaissance	1695
Shellcode	378
Worms	44

- Thống kê về dataset của nhóm:

```
# Number of features
print('Number of features: ',X.shape[1])
```

```
Number of features:  43
```

```
# Record per class
print('Record per class:\n',data.groupby('attack_cat').size())
print('\nSum:\t\t',data['attack_cat'].size)
```

```
Record per class:
 attack_cat
Analysis          2677
Backdoor          2329
DoS              16353
Exploits         44525
Fuzzers          24246
Generic          58871
Normal           93000
Reconnaissance   13987
Shellcode         1511
Worms              174
dtype: int64

Sum:             257673
```

Bot-IoT 5%

- Thống kê về dataset của bài báo:

Table 2 – Datasets summary.

	CIRA-CIC-DoHBrw-2020	Bot-IoT	UNSW NB-15
No. of features	34	29	49
No. of classes	4	5	10
No. of Samples	~1.4 million	~3 million	~0.25 million
X			

Table 6 – Attack classes in Bot-IoT dataset.

Classification of attack	No. of records	Attack name
DDoS	240,000	DDos
DoS	242,788	DoS
Reconnaissance	182,166	OS and Service Scan
Theft	160	Keylogging and Data Exfiltration

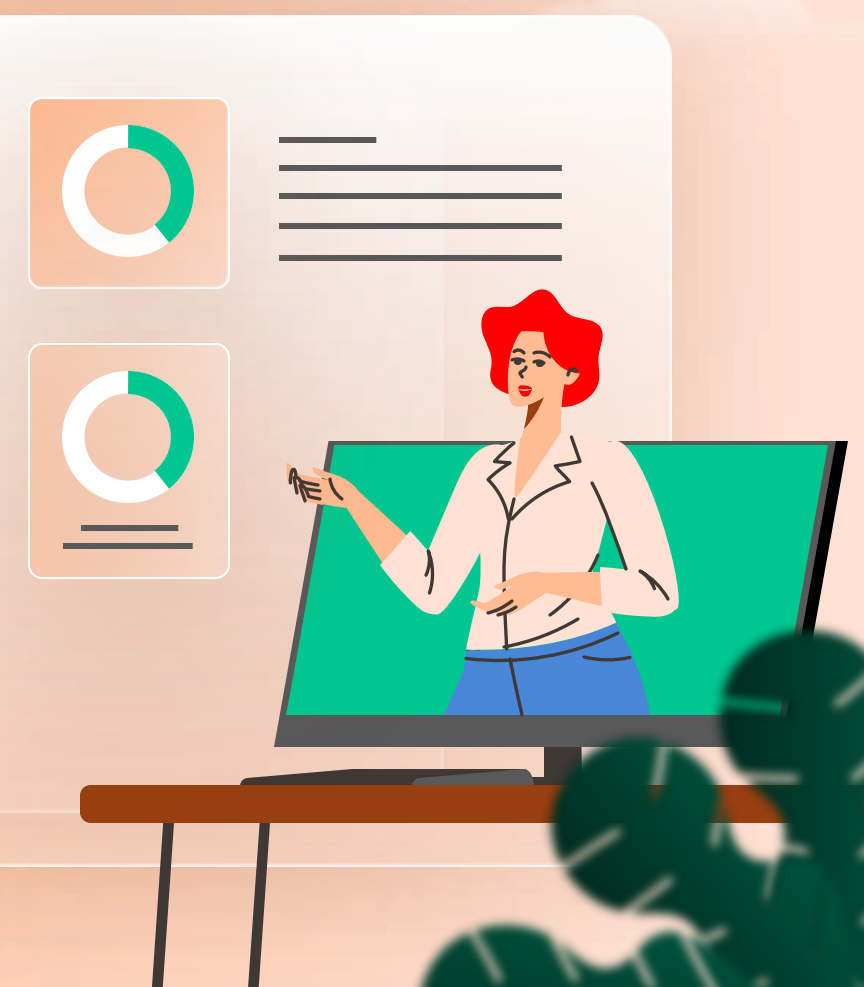
- Thống kê về dataset của nhóm:

```
[ ]  # Number of features
     print('Number of features: ',X.shape[1])

     Number of features:  43
```

```
# Record per class
print('Record per class:\n',data.groupby('category').size())
print('\nSum:\t\t',data['category'].size)

Record per class:
 category
DDoS              1926624
DoS               1650260
Normal                477
Reconnaissance      91082
Theft                  79
dtype: int64

Sum:              3668522
```

.05

Triển khai



- **Nền tảng triển khai**: Google Colab
- **Ngôn ngữ triển khai**: Python
- **Sản phẩm**: 18 model

	Full feature			**Apply GbFS**		
CIRA-CIC-DOHBrw-2020	SVM	k-NN	XgBoost	SVM	k-NN	XgBoost
UNSW-NB15	SVM	k-NN	XgBoost	SVM	k-NN	XgBoost
Bot-IoT 5%	SVM	k-NN	XgBoost	SVM	k-NN	XgBoost

- **Quy trình triển khai**:

  + *Không dùng GbFS*: Tải dataset >> tiền xử lý dữ liệu >> train model >> ghi nhận kết quả

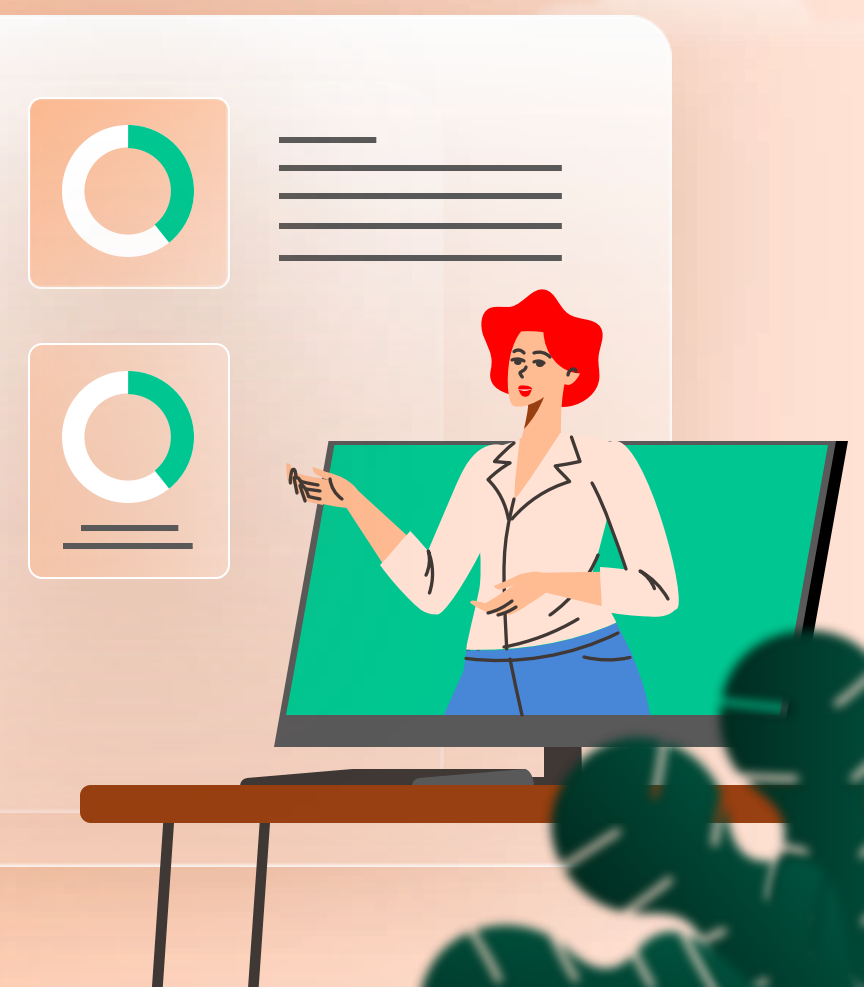
  + *Dùng GbFS*: Tải dataset >> tiền xử lý dữ liệu >> lựa chọn đặc trưng >> train model >> ghi nhận kết quả

- **Tiêu chí đánh giá của tác giả**: Độ chính xác (accuracy) và độ thu hồi (recall).

- **Tiêu chí đánh giá của nhóm**: Độ chính xác (accuracy), thời gian huấn luyện và thời gian dự đoán.

.06

Kết quả và đánh giá



CIRA-CIC-DOHBrw-2020

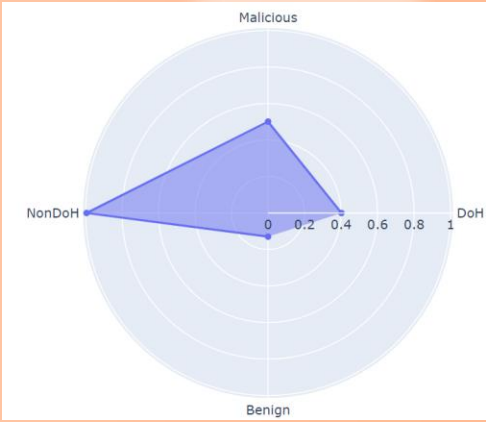
- **SVM:**

  + *Không dùng GbFS*: Độ chính xác 74%

```
## Result report
print('Train time(h):',train_time/60/60)
print('Train report:\n',classification_report(y_train, y_pred_train , zero_division=0))
print('Test report:\n',classification_report(y_test, y_pred, zero_division=0))
```

```
Train time(h): 6.203021043539048
Train report:
              precision    recall  f1-score   support

      Benign       0.83      0.88      0.86    359194
         DoH       0.44      0.02      0.05    359194
   Malicious       0.52      0.97      0.68    359194
      NonDoH       0.87      0.89      0.88    359194

    accuracy                           0.69   1436776
   macro avg       0.67      0.69      0.61   1436776
weighted avg       0.67      0.69      0.61   1436776

Test report:
              precision    recall  f1-score   support

      Benign       0.13      0.88      0.23      9827
         DoH       0.40      0.02      0.05    134623
   Malicious       0.50      0.97      0.66    125144
      NonDoH       0.99      0.89      0.94    448796

    accuracy                           0.74    718390
   macro avg       0.51      0.69      0.47    718390
weighted avg       0.79      0.74      0.71    718390
```

```
[12] start_time = time.time()
     y_pred = clf.predict(X_test)
     end_time = time.time()
     print('Predict time:', end_time - start_time, 's')
     print('Test report:\n',classification_report(y_test, y_pred, zero_division=0, digits=4))

Predict time: 531.9556746482849 s
Test report:
              precision    recall  f1-score   support

      Benign     0.1295    0.8763    0.2257      9827
         DoH     0.4018    0.0248    0.0467    134623
   Malicious     0.5009    0.9737    0.6615    125144
      NonDoH     0.9940    0.8867    0.9373    448796

    accuracy                         0.7402    718390
   macro avg     0.5066    0.6904    0.4678    718390
weighted avg     0.7853    0.7402    0.7126    718390
```

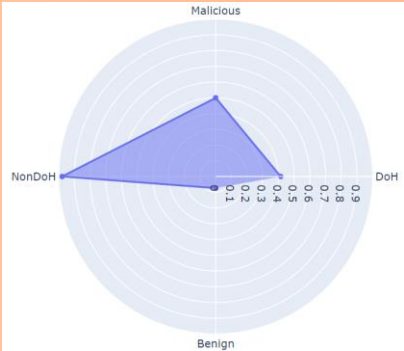


+ *Dùng GbFS*: Độ chính xác 68% với 10 đặc trưng tối ưu là ResponseTimeTimeCoefficientofVariation, DestinationIP, PacketLengthSkewFromMedian, SourcePort, ResponseTimeTimeMedian, ResponseTimeTimeSkewFromMode, PacketTimeMedian, FlowSentRate, TimeStamp, DestinationPort

```
## Result report
print('Train time(h):',train_time/60/60)
print('Train report:\n',classification_report(y_train, y_pred_train , zero_division=0))
print('Test report:\n',classification_report(y_test, y_pred, zero_division=0))
```

```
Train time(h): 0.06799053503407372
Train report:
              precision    recall  f1-score   support

      Benign       0.75      0.80      0.78    359194
         DoH       0.46      0.06      0.10    359194
   Malicious       0.52      0.92      0.66    359194
      NonDoH       0.77      0.79      0.78    359194

    accuracy                           0.64   1436776
   macro avg       0.62      0.64      0.58   1436776
weighted avg       0.62      0.64      0.58   1436776

Test report:
              precision    recall  f1-score   support

      Benign       0.07      0.80      0.14      9827
         DoH       0.41      0.05      0.10    134623
   Malicious       0.50      0.92      0.65    125144
      NonDoH       0.98      0.79      0.88    448796

    accuracy                           0.68    718390
   macro avg       0.49      0.64      0.44    718390
weighted avg       0.78      0.68      0.68    718390
```

```
start_time = time.time()
y_pred = clf.predict(X_test)
end_time = time.time()
print('Predict time:', end_time - start_time, 's')
print('Test report:\n',classification_report(y_test, y_pred, zero_division=0, digits=4))

Predict time: 0.037445783615112305 s
Test report:
              precision    recall  f1-score   support

      Benign     0.0738    0.7954    0.1351      9827
         DoH     0.4143    0.0548    0.0968    134623
   Malicious     0.4996    0.9209    0.6478    125144
      NonDoH     0.9774    0.7928    0.8755    448796

    accuracy                         0.6769    718390
   macro avg     0.4913    0.6410    0.4388    718390
weighted avg     0.7763    0.6769    0.6798    718390
```



CIRA-CIC-DOHBrw-2020

- **k-NN**:

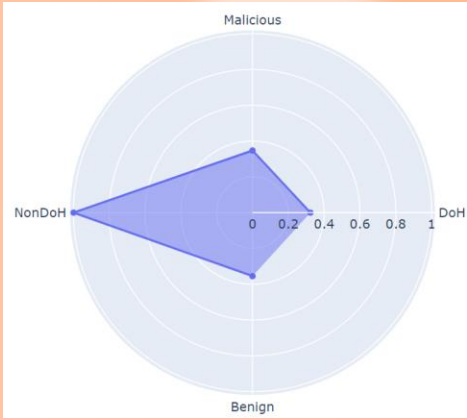
+ *Không dùng GbFS*: Độ chính xác là 75%

```
## Result report
print('Train time(h):',train_time/60/60)
print('Train report:\n',classification_report(y_train, y_pred_train , zero_division=0))
print('Test report:\n',classification_report(y_test, y_pred, zero_division=0))
```

```
Train time(h): 0.0003839625914891561
Train report:
              precision    recall  f1-score   support

      Benign       0.94      1.00      0.97    359194
         DoH       0.67      0.60      0.63    359194
   Malicious       0.68      0.70      0.69    359194
      NonDoH       1.00      0.99      1.00    359194

    accuracy                           0.82   1436776
   macro avg       0.82      0.82      0.82   1436776
weighted avg       0.82      0.82      0.82   1436776

Test report:
              precision    recall  f1-score   support

      Benign       0.35      0.77      0.49      9827
         DoH       0.32      0.29      0.31    134623
   Malicious       0.35      0.37      0.36    125144
      NonDoH       1.00      0.99      0.99    448796

    accuracy                           0.75    718390
   macro avg       0.51      0.60      0.54    718390
weighted avg       0.75      0.75      0.75    718390
```

```
start_time = time.time()
y_pred = clf.predict(X_test)
end_time = time.time()
print('Predict time:', end_time - start_time, 's')
print('Test report:\n',classification_report(y_test, y_pred, zero_division=0, digits=4))
```

```
Predict time: 1059.8100440502167 s
Test report:
              precision    recall  f1-score   support

      Benign     0.3547    0.7673    0.4851      9827
         DoH     0.3232    0.2898    0.3056    134623
   Malicious     0.3472    0.3664    0.3566    125144
      NonDoH     0.9990    0.9891    0.9940    448796

    accuracy                         0.7465    718390
   macro avg     0.5060    0.6031    0.5353    718390
weighted avg     0.7500    0.7465    0.7470    718390
```



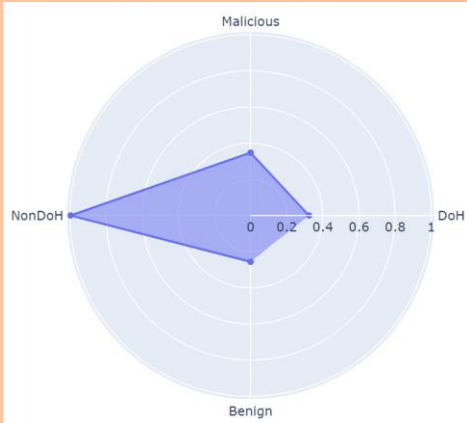
+ *Dùng GbFS*: Độ chính xác là 74%, với 10 đặc trưng tối ưu là FlowReceivedRate, PacketTimeMode, SourcePort, PacketLengthVariance, PacketTimeVariance, PacketLengthCoefficientofVariation, PacketTimeMedian, DestinationPort, PacketLengthSkewFromMedian, TimeStamp

```
## Result report
print('Train time(h):',train_time/60/60)
print('Train report:\n',classification_report(y_train, y_pred_train , zero_division=0))
print('Test report:\n',classification_report(y_test, y_pred, zero_division=0))
```

```
Train time(h): 0.0013008967373106216
Train report:
              precision    recall  f1-score   support

      Benign       0.93      1.00      0.96    359194
         DoH       0.67      0.60      0.63    359194
   Malicious       0.68      0.70      0.69    359194
      NonDoH       1.00      0.98      0.99    359194

    accuracy                           0.82   1436776
   macro avg       0.82      0.82      0.82   1436776
weighted avg       0.82      0.82      0.82   1436776

Test report:
              precision    recall  f1-score   support

      Benign       0.26      0.70      0.38      9827
         DoH       0.32      0.29      0.31    134623
   Malicious       0.35      0.37      0.36    125144
      NonDoH       1.00      0.97      0.98    448796

    accuracy                           0.74    718390
   macro avg       0.48      0.58      0.51    718390
weighted avg       0.75      0.74      0.74    718390
```

```
start_time = time.time()
y_pred = clf.predict(X_test)
end_time = time.time()
print('Predict time:', end_time - start_time, 's')
print('Test report:\n',classification_report(y_test, y_pred, zero_division=0, digits=4))
```

```
Predict time: 80.31467175483704 s
Test report:
              precision    recall  f1-score   support

      Benign     0.2559    0.7015    0.3750      9827
         DoH     0.3229    0.2900    0.3056    134623
   Malicious     0.3475    0.3666    0.3568    125144
      NonDoH     0.9962    0.9734    0.9846    448796

    accuracy                         0.7359    718390
   macro avg     0.4806    0.5829    0.5055    718390
weighted avg     0.7469    0.7359    0.7397    718390
```



CIRA-CIC-DOHBrw-2020

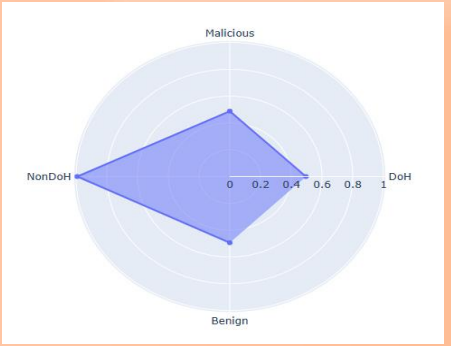
- XgBoost:
  + Không dùng GbFS: độ chính xác là 81%

Train time(h): 0.010854337149196201
Train report:

	precision	recall	f1-score	support
Benign	0.51	0.66	0.57	9980
DoH	0.50	0.76	0.61	135020
Malicious	0.51	0.21	0.29	124692
NonDoH	0.99	1.00	1.00	448697
accuracy			0.81	718389
macro avg	0.63	0.66	0.62	718389
weighted avg	0.81	0.81	0.80	718389

Test report:

	precision	recall	f1-score	support
Benign	0.49	0.64	0.56	9827
DoH	0.50	0.75	0.60	134623
Malicious	0.49	0.20	0.28	125144
NonDoH	0.99	1.00	1.00	448796
accuracy			0.81	718390
macro avg	0.62	0.65	0.61	718390
weighted avg	0.80	0.81	0.79	718390



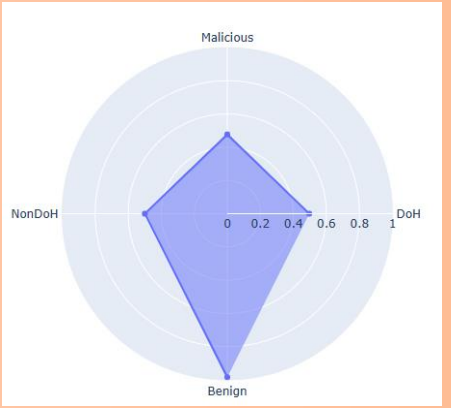
+ Dùng GbFS: độ chính xác là 80%

Train time(h): 0.0036736359861161976
Train report:

	precision	recall	f1-score	support
NonDoH	0.501514	0.713427	0.588990	9980
Malicious	0.487472	0.179107	0.261963	135020
DoH	0.499407	0.803917	0.616089	124692
Benign	0.984222	0.995549	0.989853	448697
accuracy			0.804919	718389
macro avg	0.618154	0.673000	0.614224	718389
weighted avg	0.800003	0.804919	0.782604	718389

Test report:

	precision	recall	f1-score	support
DoH	0.496189	0.708762	0.583724	9827
Malicious	0.477149	0.175423	0.256533	134623
NonDoH	0.499345	0.800997	0.615183	125144
Benign	0.983971	0.995635	0.989769	448796
accuracy			0.804101	718390
macro avg	0.614163	0.670204	0.611302	718390
weighted avg	0.797900	0.804101	0.781556	718390



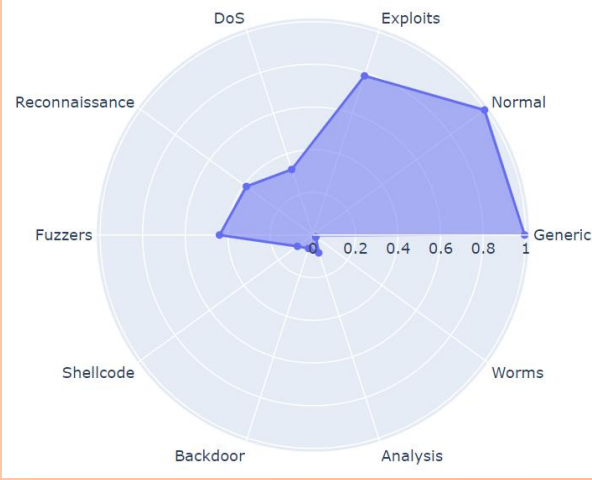
UNSW-NB15

- **SVM**:

+ *Không dùng GbFS*: Độ chính xác là 66%

```
Predict time: 1009.1137182712555 s
Test report:
              precision    recall  f1-score   support

    Analysis     0.0881    0.1659    0.1151      1320
    Backdoor     0.0663    0.4393    0.1153      1170
         DoS     0.3229    0.3875    0.3522      8148
    Exploits     0.7842    0.4045    0.5338     22378
    Fuzzers      0.4388    0.6038    0.5083     12143
    Generic      0.9945    0.9736    0.9839     29388
     Normal      0.9965    0.7032    0.8245     46435
Reconnaissance    0.3865    0.4858    0.4305      6997
  Shellcode      0.0901    0.7041    0.1598       774
      Worms      0.0169    0.8452    0.0332        84

   accuracy                         0.6641    128837
  macro avg     0.4185    0.5713    0.4057    128837
weighted avg     0.8070    0.6641    0.7111    128837
```

```
Train time(h): 1.827216714554363
Train report:
              precision    recall  f1-score   support

    Analysis     0.50      0.18      0.26      25767
    Backdoor     0.33      0.45      0.38      25767
         DoS     0.29      0.39      0.33      25767
    Exploits     0.63      0.41      0.50      25767
    Fuzzers      0.67      0.60      0.63      25767
    Generic      0.99      0.97      0.98      25767
     Normal      0.99      0.70      0.82      25767
Reconnaissance    0.38      0.50      0.43      25767
  Shellcode      0.57      0.71      0.63      25767
      Worms      0.72      0.83      0.77      25767

   accuracy                         0.57     257670
  macro avg     0.61      0.57      0.57     257670
weighted avg     0.61      0.57      0.57     257670

Test report:
              precision    recall  f1-score   support

    Analysis     0.09      0.17      0.12      1320
    Backdoor     0.07      0.44      0.12      1170
         DoS     0.32      0.39      0.35      8148
    Exploits     0.78      0.40      0.53     22378
    Fuzzers      0.44      0.60      0.51     12143
    Generic      0.99      0.97      0.98     29388
     Normal      1.00      0.70      0.82     46435
Reconnaissance    0.39      0.49      0.43      6997
  Shellcode      0.09      0.70      0.16       774
      Worms      0.02      0.85      0.03        84

   accuracy                         0.66     128837
  macro avg     0.42      0.57      0.41     128837
weighted avg     0.81      0.66      0.71     128837
```

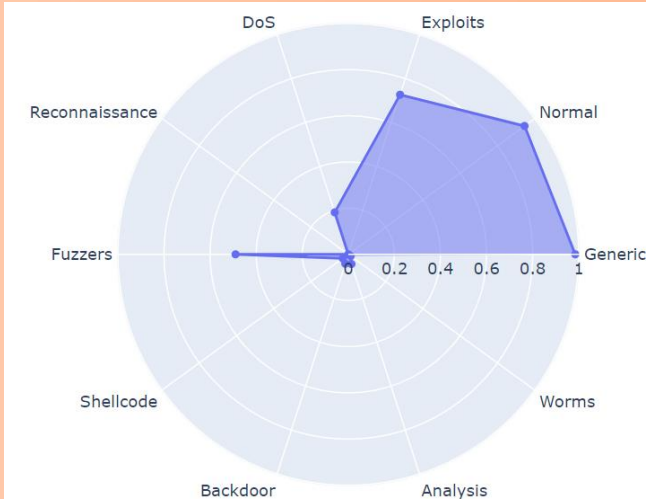


+ *Dùng GbFS*: Độ chính xác là 56% với 10 đặc trưng tối ưu là response_body_len, rate, trans_depth, ct_src_dport_ltm, ct_state_ttl, dwin, sttl, service, dttl, is_ftp_login

```
Train time(h): 0.5932960569858551
Train report:
              precision    recall  f1-score   support

    Analysis     0.22      0.01      0.01      25767
    Backdoor     0.27      0.66      0.38      25767
         DoS     0.27      0.01      0.02      25767
    Exploits     0.49      0.35      0.41      25767
    Fuzzers      0.71      0.22      0.33      25767
    Generic      0.96      0.97      0.97      25767
     Normal      0.83      0.68      0.75      25767
Reconnaissance    0.00      0.00      0.00      25767
  Shellcode      0.30      1.00      0.46      25767
      Worms      0.67      0.81      0.73      25767

   accuracy                         0.47     257670
  macro avg     0.47      0.47      0.41     257670
weighted avg     0.47      0.47      0.41     257670

Test report:
              precision    recall  f1-score   support

    Analysis     0.04      0.01      0.01      1320
    Backdoor     0.04      0.67      0.08      1170
         DoS     0.19      0.01      0.02      8148
    Exploits     0.73      0.35      0.47     22378
    Fuzzers      0.49      0.21      0.30     12143
    Generic      0.99      0.98      0.98     29388
     Normal      0.95      0.68      0.79     46435
Reconnaissance    0.00      0.00      0.00      6997
  Shellcode      0.03      1.00      0.06       774
      Worms      0.01      0.85      0.03        84

   accuracy                         0.56     128837
  macro avg     0.35      0.47      0.27     128837
weighted avg     0.75      0.56      0.62     128837
```

```
start_time = time.time()
y_pred = clf.predict(X_test)
end_time = time.time()
print('Predict time:', end_time - start_time, 's')
print('Test report:\n',classification_report(y_test, y_pred, zero_division=0, digits=4))
```

```
Predict time: 106.10473275184631 s
Test report:
              precision    recall  f1-score   support

    Analysis     0.0442    0.0061    0.0107      1320
    Backdoor     0.0429    0.6675    0.0805      1170
         DoS     0.1910    0.0114    0.0215      8148
    Exploits     0.7267    0.3497    0.4722     22378
    Fuzzers      0.4893    0.2146    0.2984     12143
    Generic      0.9851    0.9758    0.9804     29388
     Normal      0.9456    0.6806    0.7915     46435
Reconnaissance    0.0000    0.0000    0.0000      6997
  Shellcode      0.0299    0.9987    0.0581       774
      Worms      0.0130    0.8452    0.0256        84

   accuracy                         0.5623    128837
  macro avg     0.3468    0.4750    0.2739    128837
weighted avg     0.7510    0.5623    0.6216    128837
```



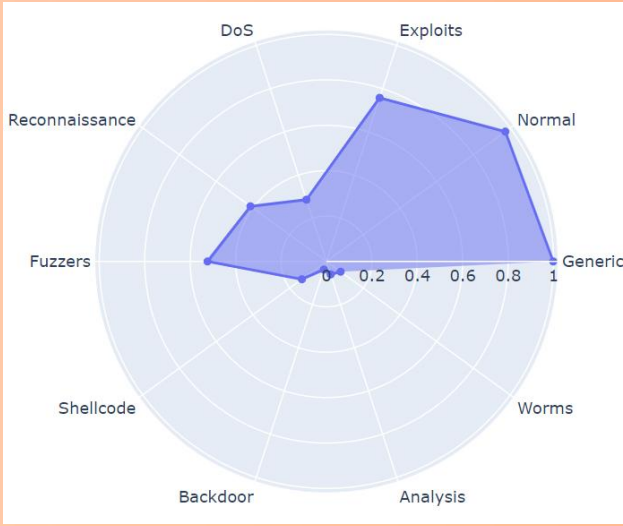
UNSW-NB15

- **k-NN**:

+ *Không dùng GbFS*: Độ chính xác là 73%

```
Predict time: 34.15016317367554 s
Test report:
                precision    recall  f1-score   support

      Analysis     0.0603    0.1894    0.0914      1320
      Backdoor     0.0375    0.1077    0.0556      1170
           DoS     0.2859    0.4256    0.3420      8148
      Exploits     0.7577    0.4562    0.5695     22378
      Fuzzers     0.5234    0.6605    0.5840     12143
      Generic     0.9983    0.9738    0.9859     29388
       Normal     0.9733    0.8280    0.8948     46435
Reconnaissance     0.4132    0.6025    0.4902      6997
    Shellcode     0.1336    0.2997    0.1848       774
        Worms     0.0772    0.2619    0.1192        84

     accuracy                         0.7266    128837
    macro avg     0.4260    0.4805    0.4317    128837
 weighted avg     0.8018    0.7266    0.7522    128837
```

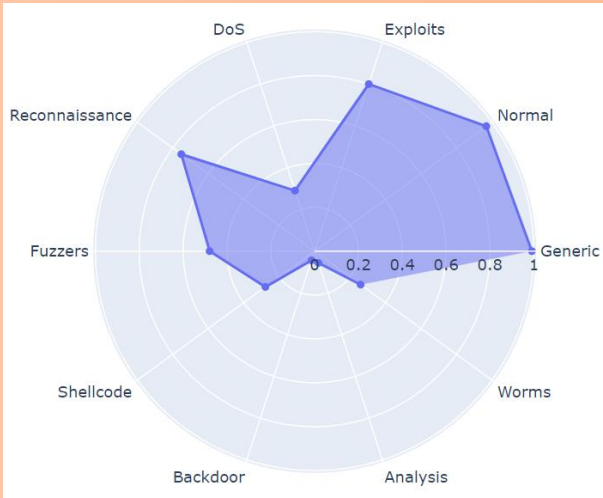
```
Train time(h): 6.594949298434787e-05
Train report:
                precision    recall  f1-score   support

      Analysis     0.85      1.00      0.92     25767
      Backdoor     0.86      1.00      0.93     25767
           DoS     0.67      0.69      0.68     25767
      Exploits     0.79      0.54      0.64     25767
      Fuzzers     0.80      0.77      0.79     25767
      Generic     1.00      0.97      0.99     25767
       Normal     0.95      0.86      0.90     25767
Reconnaissance     0.79      0.80      0.79     25767
    Shellcode     0.93      1.00      0.97     25767
        Worms     0.99      1.00      0.99     25767

     accuracy                         0.86    257670
    macro avg     0.86      0.86      0.86    257670
 weighted avg     0.86      0.86      0.86    257670

Test report:
                precision    recall  f1-score   support

      Analysis     0.06      0.19      0.09      1320
      Backdoor     0.04      0.11      0.06      1170
           DoS     0.29      0.43      0.34      8148
      Exploits     0.76      0.46      0.57     22378
      Fuzzers     0.52      0.66      0.58     12143
      Generic     1.00      0.97      0.99     29388
       Normal     0.97      0.83      0.89     46435
Reconnaissance     0.41      0.60      0.49      6997
    Shellcode     0.13      0.30      0.18       774
        Worms     0.08      0.26      0.12        84

     accuracy                         0.73    128837
    macro avg     0.43      0.48      0.43    128837
 weighted avg     0.80      0.73      0.75    128837
```



+ *Dùng GbFS*: Độ chính xác là 72% với 10 đặc trưng tối ưu là smean, ct_src_ltm, dload, ct_state_ttl, dttl, sjit, ct_srv_dst, dwin, sttl, sinpkt

```
Train time(h): 0.0002135569519466824
Train report:
                precision    recall  f1-score   support

      Analysis     0.45      0.65      0.53     25767
      Backdoor     0.49      0.60      0.54     25767
           DoS     0.44      0.38      0.41     25767
      Exploits     0.76      0.52      0.62     25767
      Fuzzers     0.78      0.79      0.78     25767
      Generic     0.99      0.97      0.98     25767
       Normal     0.94      0.80      0.86     25767
Reconnaissance     0.93      0.81      0.86     25767
    Shellcode     0.95      1.00      0.98     25767
        Worms     0.99      1.00      1.00     25767

     accuracy                         0.75    257670
    macro avg     0.77      0.75      0.76    257670
 weighted avg     0.77      0.75      0.76    257670

Test report:
                precision    recall  f1-score   support

      Analysis     0.06      0.35      0.10      1320
      Backdoor     0.04      0.24      0.07      1170
           DoS     0.29      0.29      0.29      8148
      Exploits     0.80      0.47      0.59     22378
      Fuzzers     0.48      0.71      0.57     12143
      Generic     0.99      0.97      0.98     29388
       Normal     0.97      0.77      0.86     46435
Reconnaissance     0.75      0.76      0.76      6997
    Shellcode     0.28      0.59      0.38       774
        Worms     0.26      0.80      0.39        84

     accuracy                         0.72    128837
    macro avg     0.49      0.60      0.50    128837
 weighted avg     0.82      0.72      0.75    128837
```

```
Predict time: 11.20298457145691 s
Test report:
                precision    recall  f1-score   support

      Analysis     0.0565    0.3545    0.0974      1320
      Backdoor     0.0440    0.2359    0.0742      1170
           DoS     0.2900    0.2921    0.2910      8148
      Exploits     0.8003    0.4727    0.5944     22378
      Fuzzers     0.4789    0.7117    0.5725     12143
      Generic     0.9913    0.9720    0.9815     29388
       Normal     0.9684    0.7718    0.8590     46435
Reconnaissance     0.7516    0.7606    0.7561      6997
    Shellcode     0.2774    0.5891    0.3772       774
        Worms     0.2597    0.7976    0.3918        84

     accuracy                         0.7187    128837
    macro avg     0.4918    0.5958    0.4995    128837
 weighted avg     0.8213    0.7187    0.7544    128837
```



UNSW-NB15

- XgBoost:

+ Không dùng GbFS: độ chính xác 84%

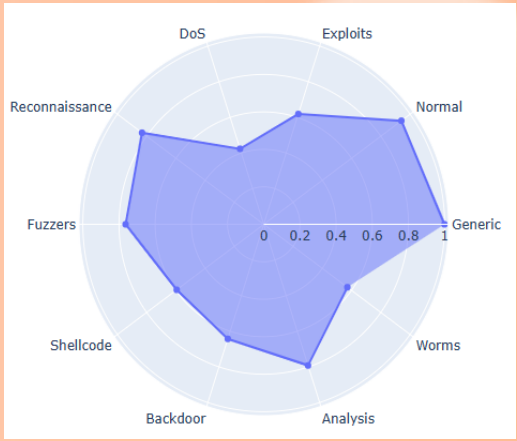
Train time(h): 0.0013035198052724203

Train report:

	precision	recall	f1-score	support
Analysis	0.85	0.05	0.10	1357
Backdoor	0.71	0.15	0.25	1159
DoS	0.44	0.20	0.27	8205
Exploits	0.62	0.87	0.72	22147
Fuzzers	0.77	0.68	0.72	12103
Generic	1.00	0.98	0.99	29483
Normal	0.94	0.96	0.95	46565
Reconnaissance	0.84	0.74	0.79	6990
Shellcode	0.64	0.38	0.48	737
Worms	0.72	0.56	0.63	90
accuracy			0.84	128836
macro avg	0.75	0.56	0.59	128836
weighted avg	0.84	0.84	0.83	128836

Test report:

	precision	recall	f1-score	support
Analysis	0.79	0.05	0.09	1320
Backdoor	0.64	0.15	0.24	1170
DoS	0.42	0.20	0.27	8148
Exploits	0.62	0.86	0.72	22378
Fuzzers	0.76	0.68	0.72	12143
Generic	1.00	0.97	0.99	29388
Normal	0.94	0.96	0.95	46435
Reconnaissance	0.83	0.73	0.78	6997
Shellcode	0.59	0.35	0.44	774
Worms	0.57	0.43	0.49	84
accuracy			0.84	128837
macro avg	0.72	0.54	0.57	128837
weighted avg	0.84	0.84	0.83	128837



+ Dùng GbFS: độ chính xác 82%

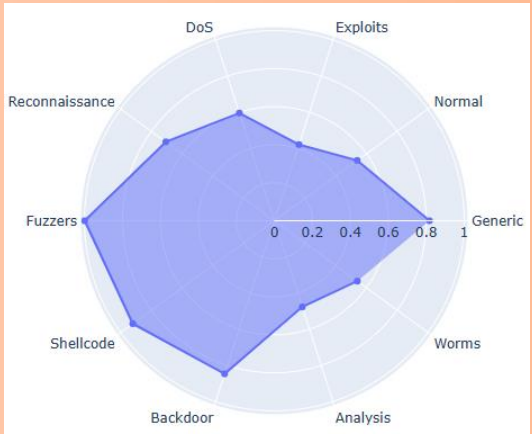
Train time(h): 0.002300196157561408

Train report:

	precision	recall	f1-score	support
Generic	0.815534	0.061901	0.115068	1357
Shellcode	0.633333	0.114754	0.194302	1159
Reconnaissance	0.441341	0.163681	0.238798	8205
Normal	0.596109	0.837089	0.696340	22147
Backdoor	0.714537	0.636784	0.673424	12103
Exploits	0.999303	0.973035	0.985994	29483
Fuzzers	0.922162	0.959433	0.940429	46565
DoS	0.846006	0.754506	0.797641	6990
Analysis	0.552941	0.127544	0.207277	737
Worms	0.542857	0.211111	0.304000	90
accuracy			0.827075	128836
macro avg	0.706412	0.483984	0.515327	128836
weighted avg	0.823410	0.827075	0.811339	128836

Test report:

	precision	recall	f1-score	support
Generic	0.817308	0.064394	0.119382	1320
Normal	0.539906	0.098291	0.166305	1170
Exploits	0.421274	0.159917	0.231830	8148
DoS	0.595799	0.829029	0.693325	22378
Reconnaissance	0.706399	0.635428	0.669037	12143
Fuzzers	0.998952	0.973152	0.985883	29388
Shellcode	0.921268	0.958329	0.939433	46435
Backdoor	0.846712	0.750750	0.795849	6997
Analysis	0.476684	0.118863	0.190279	774
Worms	0.540541	0.238095	0.330579	84
accuracy			0.824569	128837
macro avg	0.686484	0.482625	0.512190	128837
weighted avg	0.819087	0.824569	0.808927	128837



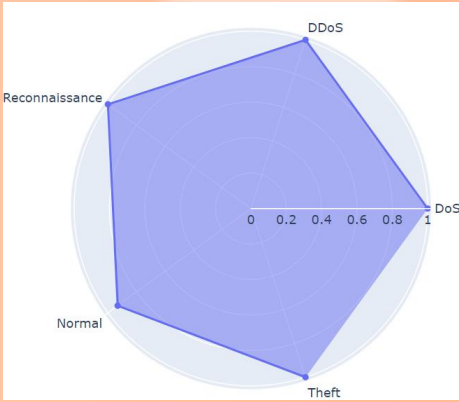
Bot-IoT 5%

- **SVM**:

+ *Không dùng GbFS*: Độ chính xác là 99.9989%

```
Predict time: 36.45159840583801 s
Test report:
                precision    recall  f1-score   support

          DDoS   1.000000  1.000000  1.000000    963821
           DoS   1.000000  1.000000  1.000000    824747
        Normal   0.932000  0.987288  0.958848       236
Reconnaissance   0.999934  0.999626  0.999780     45416
         Theft   1.000000  1.000000  1.000000        41

      accuracy                       0.999989   1834261
     macro avg   0.986387  0.997383  0.991726   1834261
  weighted avg   0.999990  0.999989  0.999989   1834261
```

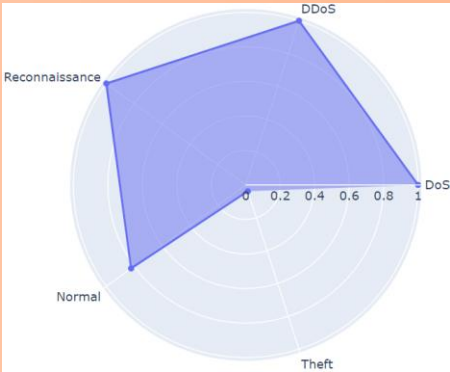
```
Train time(h): 4.122527541716893
Train report:
                precision    recall  f1-score   support

          DDoS   1.000000  1.000000  1.000000    733704
           DoS   1.000000  1.000000  1.000000    733704
        Normal   0.999847  1.000000  0.999924    733704
Reconnaissance   1.000000  0.999847  0.999924    733704
         Theft   1.000000  1.000000  1.000000    733704

      accuracy                       0.999969   3668520
     macro avg   0.999969  0.999969  0.999969   3668520
  weighted avg   0.999969  0.999969  0.999969   3668520

Test report:
                precision    recall  f1-score   support

          DDoS   1.000000  1.000000  1.000000    963821
           DoS   1.000000  1.000000  1.000000    824747
        Normal   0.932000  0.987288  0.958848       236
Reconnaissance   0.999934  0.999626  0.999780     45416
         Theft   1.000000  1.000000  1.000000        41

      accuracy                       0.999989   1834261
     macro avg   0.986387  0.997383  0.991726   1834261
  weighted avg   0.999990  0.999989  0.999989   1834261
```



+ *Dùng GbFS*: Độ chính xác là 99.93% với 10 đặc trưng tối ưu là AR_P_Proto_P_Dport, saddr, N_IN_Conn_P_DstIP, stime, pkSeqID, state, sum, flgs, daddr, bytes

```
Train time(h): 0.03805756919913821
Train report:
                precision    recall  f1-score   support

          DDoS   1.000000  0.998878  0.999439    733704
           DoS   0.999936  1.000000  0.999968    733704
        Normal   0.998884  1.000000  0.999442    733704
Reconnaissance   1.000000  0.998752  0.999375    733704
         Theft   0.998813  1.000000  0.999406    733704

      accuracy                       0.999526   3668520
     macro avg   0.999526  0.999526  0.999526   3668520
  weighted avg   0.999526  0.999526  0.999526   3668520

Test report:
                precision    recall  f1-score   support

          DDoS   1.000000  0.998875  0.999437    963821
           DoS   0.999899  1.000000  0.999950    824747
        Normal   0.822300  1.000000  0.902486       236
Reconnaissance   1.000000  0.998503  0.999251     45416
         Theft   0.038716  1.000000  0.074545        41

      accuracy                       0.999372   1834261
     macro avg   0.772183  0.999476  0.795134   1834261
  weighted avg   0.999910  0.999372  0.999630   1834261
```

```
Predict time: 0.09997320175170898 s
Test report:
                precision    recall  f1-score   support

          DDoS   1.000000  0.998875  0.999437    963821
           DoS   0.999899  1.000000  0.999950    824747
        Normal   0.822300  1.000000  0.902486       236
Reconnaissance   1.000000  0.998503  0.999251     45416
         Theft   0.038716  1.000000  0.074545        41

      accuracy                       0.999372   1834261
     macro avg   0.772183  0.999476  0.795134   1834261
  weighted avg   0.999910  0.999372  0.999630   1834261
```

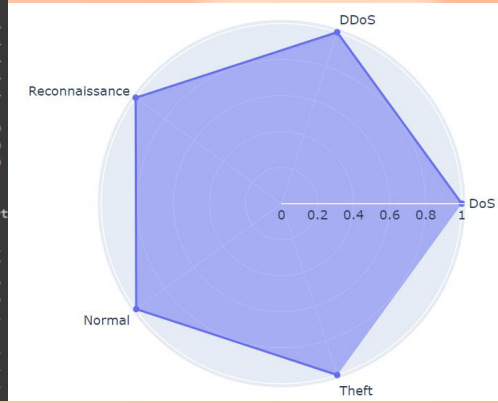


Bot-IoT 5%

- **k-NN**:
  + *Không dùng GbFS*: Độ chính xác là 99.9983%

```
Predict time: 11817.05957365036 s
Test report:
                precision    recall  f1-score   support

          DDoS   0.999992  0.999980  0.999986    963821
           DoS   0.999978  0.999990  0.999984    824747
        Normal   0.995726  0.987288  0.991489       236
Reconnaissance   0.999912  0.999978  0.999945     45416
         Theft   1.000000  1.000000  1.000000        41

      accuracy                       0.999983   1834261
     macro avg   0.999122  0.997447  0.998281   1834261
  weighted avg   0.999983  0.999983  0.999983   1834261
```

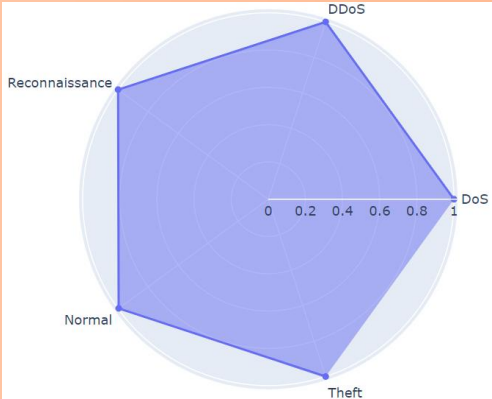
```
Train time(h): 0.0011156907346513536
Train report:
                precision    recall  f1-score   support

          DDoS       1.00      1.00      1.00    733704
           DoS       1.00      1.00      1.00    733704
        Normal       1.00      1.00      1.00    733704
Reconnaissance       1.00      1.00      1.00    733704
         Theft       1.00      1.00      1.00    733704

      accuracy                           1.00   3668520
     macro avg       1.00      1.00      1.00   3668520
  weighted avg       1.00      1.00      1.00   3668520

Test report:
                precision    recall  f1-score   support

          DDoS       1.00      1.00      1.00    963821
           DoS       1.00      1.00      1.00    824747
        Normal       1.00      0.99      0.99       236
Reconnaissance       1.00      1.00      1.00     45416
         Theft       1.00      1.00      1.00        41

      accuracy                           1.00   1834261
     macro avg       1.00      1.00      1.00   1834261
  weighted avg       1.00      1.00      1.00   1834261
```



+ Dùng GbFS: Độ chính xác là 99.9993% với 10 đặc trưng tối ưu là AR_P_Proto_P_Dport, saddr, N_IN_Conn_P_DstIP, stime, pkSeqID, state, sum, flgs, daddr, bytes

```
Train time(h): 0.005188713404867385
Train report:
                precision    recall  f1-score   support

          DDoS   0.999997  0.999993  0.999995    733704
           DoS   0.999996  0.999997  0.999997    733704
        Normal   1.000000  1.000000  1.000000    733704
Reconnaissance   0.999997  1.000000  0.999999    733704
         Theft   1.000000  1.000000  1.000000    733704

      accuracy                       0.999998   3668520
     macro avg   0.999998  0.999998  0.999998   3668520
  weighted avg   0.999998  0.999998  0.999998   3668520

Test report:
                precision    recall  f1-score   support

          DDoS   0.999999  0.999992  0.999995    963821
           DoS   0.999993  1.000000  0.999996    824747
        Normal   0.995726  0.987288  0.991489       236
Reconnaissance   0.999890  0.999978  0.999934     45416
         Theft   1.000000  0.975610  0.987654        41

      accuracy                       0.999993   1834261
     macro avg   0.999122  0.992574  0.995814   1834261
  weighted avg   0.999993  0.999993  0.999993   1834261
```

```
Predict time: 265.37616872787476 s
Test report:
                precision    recall  f1-score   support

          DDoS   0.999999  0.999992  0.999995    963821
           DoS   0.999993  1.000000  0.999996    824747
        Normal   0.995726  0.987288  0.991489       236
Reconnaissance   0.999890  0.999978  0.999934     45416
         Theft   1.000000  0.975610  0.987654        41

      accuracy                       0.999993   1834261
     macro avg   0.999122  0.992574  0.995814   1834261
  weighted avg   0.999993  0.999993  0.999993   1834261
```



Bot-IoT 5%

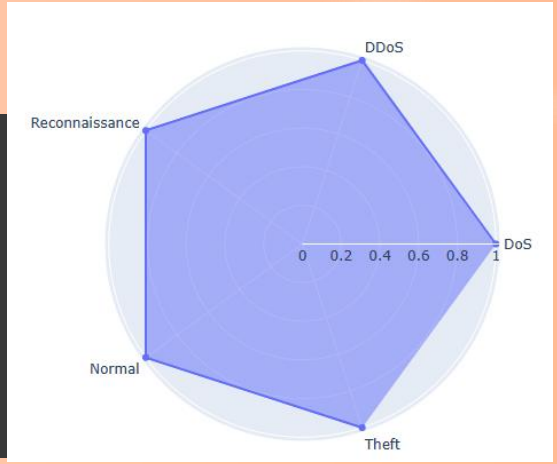
- XgBoost:
  + Không dùng GbFS: độ chính xác 99%

Train time(h): 0.01582732684082455
Train report:

	precision	recall	f1-score	support
DoS	1.000000	1.000000	1.000000	962803
DDoS	1.000000	1.000000	1.000000	825513
Reconnaissance	1.000000	0.975104	0.987395	241
Normal	0.999869	1.000000	0.999934	45666
Theft	1.000000	1.000000	1.000000	38
accuracy			0.999997	1834261
macro avg	0.999974	0.995021	0.997466	1834261
weighted avg	0.999997	0.999997	0.999997	1834261

Test report:

	precision	recall	f1-score	support
DoS	0.999999	1.000000	0.999999	963821
DDoS	1.000000	1.000000	1.000000	824747
Reconnaissance	1.000000	0.957627	0.978355	236
Normal	0.999780	1.000000	0.999890	45416
Theft	1.000000	0.975610	0.987654	41
accuracy			0.999994	1834261
macro avg	0.999956	0.986647	0.993180	1834261
weighted avg	0.999994	0.999994	0.999994	1834261



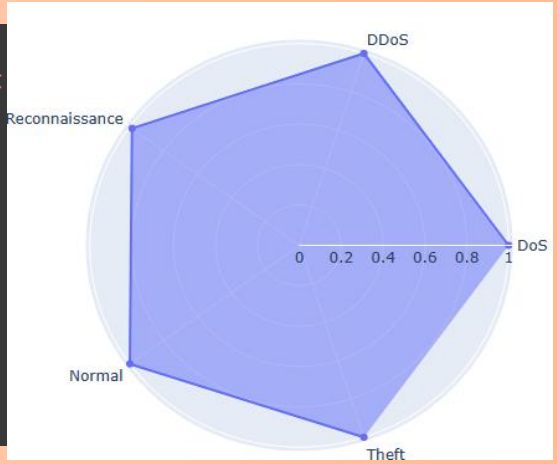
  + Dùng GbFS: độ chính xác 99%

Train time(h): 0.005960666007465786
Train report:

	precision	recall	f1-score	support
DoS	0.999970	1.000000	0.999985	962803
DDoS	1.000000	1.000000	1.000000	825513
Reconnaissance	1.000000	0.946058	0.972281	241
Normal	0.999869	0.999540	0.999704	45666
Theft	1.000000	0.973684	0.986667	38
accuracy			0.999981	1834261
macro avg	0.999968	0.983856	0.991727	1834261
weighted avg	0.999981	0.999981	0.999981	1834261

Test report:

	precision	recall	f1-score	support
DoS	0.999962	1.000000	0.999981	963821
DDoS	1.000000	1.000000	1.000000	824747
Reconnaissance	0.986364	0.919492	0.951754	236
Normal	0.999758	0.999361	0.999560	45416
Theft	1.000000	0.926829	0.962025	41
accuracy			0.999972	1834261
macro avg	0.997217	0.969136	0.982664	1834261
weighted avg	0.999972	0.999972	0.999972	1834261



Kết luận chung

Về mặt lý thuyết của giải pháp được đề xuất

Xét về lý thuyết, đây là 1 ý tưởng đúng khi giảm số đặc trưng là các đặc trưng gây ảnh xấu đến hiệu suất thuật toán học để tăng hiệu suất cho bộ phân loại.

Lý tưởng thì GbFs sẽ tìm ra bộ các đặc trưng mà thỏa mãn 2 điều kiện là mang lại hiệu suất học cao (A_i) và có độ tương quan thấp hay có độ đa dạng dữ liệu cao ($1 - M_i$).

Các đặc trưng tối ưu tìm được sẽ mang lại các lợi ích:

- Giảm chiều dataset -> Giảm thời gian huấn luyện và dự đoán của mô hình
- Loại bỏ các đặc trưng kém đa dạng (tương quan cao), không có ý nghĩa trong việc phân loại; chọn ra các đặc trưng đa dạng ($1-M_i$) và mang lại hiệu suất cao (A_i) -> Tăng hiệu suất/độ chính xác cho mô hình học

Kết luận chung

Về mặt thực tiễn của giải pháp được đề xuất

Dựa trên thí nghiệm thực tế, kết quả thu được đã **không thể chứng minh** rằng GbFS sẽ chắc chắn tăng độ chính xác cho mô hình học. **Tuy nhiên**, cũng chứng minh rằng GbFS đã giúp cho các mô hình học giảm thời huấn luyện và dự đoán, điều này mang giá trị thực tiễn cao, do trong thực tế có rất nhiều gói tin được gửi đến trong 1 thời gian ngắn nên mô hình không chỉ cần độ chính xác mà còn cần tốc độ dự đoán.

Thông qua đánh giá một cách "con người" thì nhóm cho rằng các đặc trưng mà GbFS chọn ra có các đặc trưng không có ý nghĩa trong việc phân loại (ví dụ: timestamp, …); nguyên nhân cho việc này là do thiết kế của hàm fitness mà nhóm tác giả đề xuất chưa hợp lý (đa dạng dữ liệu có thể mang lại hiệu suất học cao nhưng không có nghĩa là có ý nghĩa trong việc phân loại).

Kết luận chung

Đánh giá của nhóm

GbFS do bài báo đề xuất về mặt thuật toán thì rất tốt, thuật toán liên tục tìm bộ đặc trưng "phù hợp" để mang lại hiệu suất cao cho mô hình học thông qua việc tính giá trị fitness cho từng NST sau mỗi giai đoạn tạo ra quần thể mới.

Tuy nhiên, điểm quan trọng của thuật toán cũng là điểm yếu của nó, chính là hàm fitness. Hàm fitness là hàm dùng tính giá trị fitness, mà giá trị fitness ở đây dùng để thể hiện độ ảnh hưởng của NST đến hiệu suất của mô hình học và độ đa dạng của đặc trưng. Dễ dàng thấy được rằng việc đặc trưng đó độ đa dạng cao không có nghĩa là nó có ý nghĩa trong việc phân loại các gói tin (ví dụ: timestamp, …).

Tuy GbFS do nhóm thiết kế **không thể chắc chắn gia tăng độ chính xác** cho mô hình học và cũng không chắc chắn GbFS của nhóm giống với của tác giả, nhưng cũng đã **giảm đáng kể thời gian huấn luyện và dự đoán** của mô hình, điều này rất có ý nghĩa trong thực tiễn.

THANKS FOR WATCHING