Explainability of Machine Learning in Intrusion Detection Systems

Motivation

- Increasing Network Security Threat
- Huge economic loss due to attacks

Potential problems of machine learning models in IDS (Intrusion Detection Systems)

- Semantic Gap
- Shortcut Learning
- High Cost of Errors

Definition of Explainability

- The degree to which a human can understand the cause of a decision

Data Preprocessing

- Remove features with only 1 unique value
- Drop records with Nan Values
- Build reference table for string types
  - Find unique value of each features
  - Build a table based on the unique value
- Replace the original string values with index on reference table

Algorithm 1 Global Explanation

1: procedure PARABOLIC($r, D, I, G, E, S$)
2: Begin Procedure:
3: Initialize dataset using black-box
4: $D = \pi(\text{Weird Data})$
5: Initialize default parameter grids
6: $G = G_0$
7: Initialize temporary grids
8: $G_{temp} = \text{None}$
9: for $i \leftarrow 1 \text{ to } 3$ do
10: Use grid-search on the dataset using grid $G$
11: $G' \leftarrow \text{Grid} - \text{Search}(D, G)$
12: if $G' \neq G_{temp}$ then
13: Break
14: end if
15: Update temporary grid $G_{temp}$
16: $G_{temp} \leftarrow G'$
17: Update grid that centered at $G'$
18: $G \leftarrow \text{Matrix}(G)$
19: end for
20: Get the best model $G_{best}$
21: $G_{best} \leftarrow G'$
22: Initialize Explanation $E$
23: $E \leftarrow G_{best}$
24: for $i \leftarrow 1 \text{ to } I$ do
25: Sample dataset $D'$
26: $D' = \text{Split}(D, S)$
27: Train the model $G_{best}$
28: $G' \leftarrow \text{Fit}(G_{best}, D')$
29: if $G_{score} > G_{score}$ then
30: $E \leftarrow G'$
31: end if
32: end for
33: Prune selected tree $E \leftarrow \text{Prune}(E)$
34: return $E$
35: end procedure

Conclusion

- For maliciously predicted traffic, tell the network operators about the features and corresponding values that might cause the prediction result based on the distribution
- Provide a improved formula as the firewall rule based on the data

Future Work

- Self-update the formula for the firewall rule based on the new incoming traffics
- Implement the user interface