THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Presentation #22306 for:
National Defense Industrial Association

22nd Annual Systems and Mission Engineering Conference
October 21 – 24 2019, Tampa, FL

Ying Zhou, PhD student
Dr. Thomas A. Mazzuchi, Professor
Dr. Shahram Sarkani, Professor

Department of Systems Engineering
George Washington University

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Background:

- While machine learning has demonstrated a great potential in network intrusion detection, most machine Learning systems especially deep learning methods are often regarded as black boxes which are hard to explain.
- Explainable machine learning systems are expected to offer rationales for classification decisions which are consistent with domain expert knowledge.
- For instance, subject matter experts consider that a high number of connections with low duration and low login success rate are suspicious and match the characteristics of certain network intrusion attacks.
- Explainable machine learning systems are essential for cyber security professionals to understand, trust, and implement the network attack classifications.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Typical Research Questions:

- What are the most relevant features to improve classifier performance for detecting network intrusion?
- Does a certain combination of feature selection method and classification algorithm perform better than the other?
- As more complex machine learning systems are often required to improve accuracy, how do researchers balance the tradeoff between interpretability and accuracy?

My Research Objective:

My research will apply new approaches for constructing machine learning systems with both high performance and interpretability to better detect network intrusions.

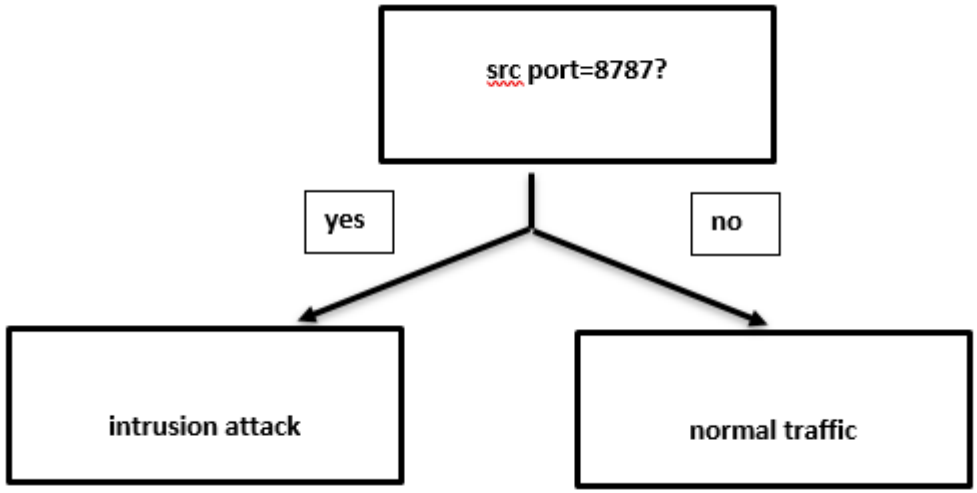
Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Black Box Machine learning vs Explainable Machine learning

- Black Box Machine learning: Only focus on model performance for correctly classifying the instances, generally apply deep learning models with complex structure, slow computation and precise predictive performance, difficult to explain how model made the decision to general audience, Example: Artificial Neural Network

- Explainable Machine learning: Apply machine learning algorithms for classification and meanwhile explain model decisions with various methods

An illustrative example: machine learning models considered interpretable due to their simple structure such as decision trees and rule lists



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

The rule for the portsweep attack:

IF wrong_fragment <=0 AND num_compromised <=0 AND count <=2 AND dst_host_srv_diff_host_ rate <=0.24 AND dst_host_same_srv_rate <=0.01 AND src_bytes <=1 AND rerror_rate <=0.98 AND serror_rate <=0.32 AND protocol_type=tcp THEN attack=portsweep

The rule for the satan attack:

IF wrong_fragment <=0 AND num_compromised <=0 AND count <=2 AND dst_host_srv_diff_host_ rate <=0.24 AND dst_host_same_srv_rate <=0.01 AND src_bytes <=1 AND rerror_rate <=0.98 AND serror_rate <=0.32 AND protocol_type=udp THEN attack=satan

Reference:An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks, Expert Systems with Applications · November 2005

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Example: application of interpretation methods to machine learning algorithms after model training

LIME & SHAP both result in a feature contribution score, s

s > 0: the feature pushes the classifier toward attack class

s < 0: the feature pushes the classifier toward normal class

The closer s is to 0, the weaker the feature's contribution

An illustrative example: Top five features contribute the most to the malware classification

Feature 367
Group: header file info
SHAP:1.78
Feature 650
Group: general file info
SHAP:0.45
Feature 30
Group: header file info
SHAP:0.42
Feature 570
Group Byte Histogram
SHAP -0.33
Feature 608
Group general file info
SHAP 0.31

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Summary of Types of Interpretability Techniques:

- Global interpretability facilitates understanding the entire logic of a model for classifying different possible outcomes, example: decision tree/ rule lists
- local interpretability focuses on justifying why the model made a specific decision for an instance, example: LIME & SHAP

- Model specific interpretability methods are limited to specific algorithms, example: decision tree/ rule lists
- Model agnostic methods can be applied to any types of machine learning algorithms, example: LIME & SHAP

- Intrinsic interpretability refers to machine learning models that are considered interpretable due to their simple structure, example: decision tree/ rule lists
- Post hoc interpretability refers to the application of interpretation methods after model training, example: LIME & SHAP

Reference:Interpretable Machine Learning, *A Guide for Making Black Box Models Explainable*

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Typical Data Characteristics in Network Intrusion Detection:

- large scale data with multiple classes of attacks
- Data imbalance, #attacks<<#normal traffic, biased classifier
  - RUS: Random Undersampling
  - ROS: Random Oversampling
  - SMOTE: Synthetic Minority Over-sampling Technique
- High dimensionality
- Highly correlated features
- Irrelevant features/misleading features

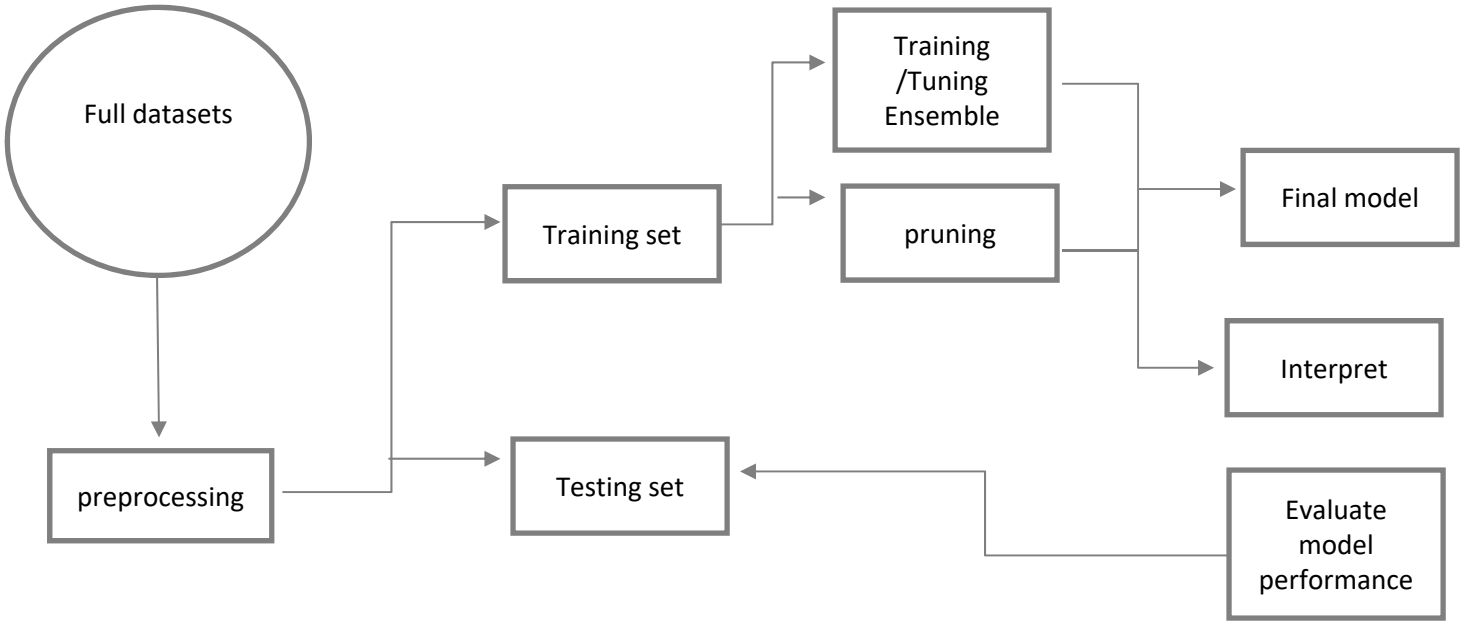
Example NSL-KDD Dataset:

Traffic Class	Count
Normal	67343 (53.46%)
DoS	45927 (36.46%)
Probe	11656 (9.25%)
R2L	995 (0.79%)
U2R	52 (0.04%)

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Machine Learning System Flow Chart



Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Introduction of Multivariate Adaptive Regression Splines (MARS)

- Multivariate adaptive regression splines (MARS) is a multivariate nonlinear and nonparametric classification/regression technique.
- The MARS model is built by fitting basis functions to distinct intervals of the predictors. The relationship between the response variable and the predictor variables is discovered from a set of basis functions and their coefficients.
- The variables to be used and the knot points of the intervals for each predictor are found through a quick but intensive search process.

Reference:Multivariate Adaptive Regression Splines, Ann. Statist. Volume 19, Number 1 (1991), 1-67.

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Advantage of MARS

- MARS is capable of modeling complex non-linear relationship among variables without imposing strong modeling assumptions.

- MARS can capture the relative importance of the independent variables to the dependent variable when a lot of potential independent variables are considered.

- MARS does not need time consuming training process and hence can save lots of model building time, especially when the dataset is large scale.

- MARS can be easily interpreted. It can point out which variables are important in classifying observations, meanwhile it also indicates a particular instance belongs to a specific class when the modeling rules are met.

Reference:Multivariate adaptive regression splines and neural network models for prediction of pile drivability,Geoscience Frontiers

General Form of MARS Model

$$\widehat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$$

The model is a weighted sum of basis functions $B_i(x)$. Each c_i is a constant coefficient.

$B_i(x)$ takes one of the following three forms:

1) a constant 1.

2) a *hinge* function.

A hinge function has the form $\max(0, x - \text{constant})$ or $\max(0, \text{constant} - x)$.

3) a product of two or more hinge functions.

Reference:Multivariate Adaptive Regression Splines, Ann. Statist. Volume 19, Number 1 (1991), 1-67.

MARS Fitting Process

The optimal MARS model is selected in a two-step procedure

- In the first stage, MARS constructs a very large number of basis functions to overfit the data initially. An overfit model has a good fit to the training data but will not generalize well to the test data.

- In the second stage, basis functions are removed in the order of least contributions using the generalized cross-validation (GCV) statistics. Then a measure of variable importance can be estimated by observing the decrease in the calculated GCV when a variable is removed from the model. This pruning process will continue until the remaining basis functions all satisfying the pre-defined criteria.

Reference:Multivariate Adaptive Regression Splines, Ann. Statist. Volume 19, Number 1 (1991), 1-67.

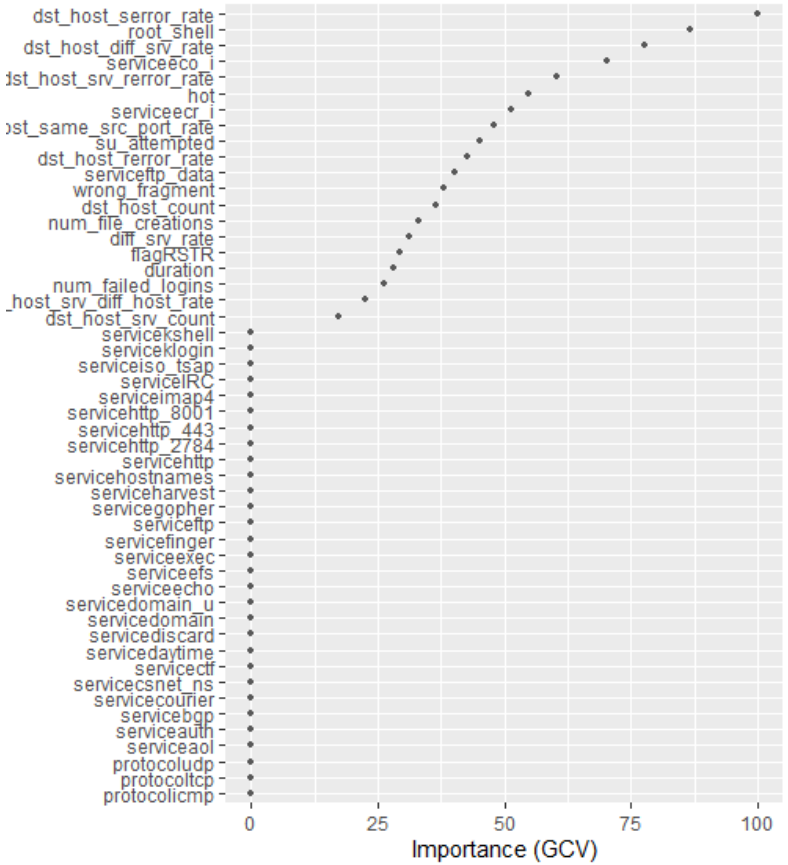
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Preliminary MARS Model Results

120 features are reduced to 20

```
(Intercept)
serviceeco_i
serviceecr_i
serviceftp_data
flagRSTR
root_shell
su_attempted
h(1326-duration)
h(duration-1326)
h(1-wrong_fragment)
h(wrong_fragment-1)
h(3-hot)
h(hot-3)
h(1-num_failed_logins)
h(num_failed_logins-1)
h(4-num_file_creations)
h(num_file_creations-4)
h(0.05-diff_srv_rate)
h(diff_srv_rate-0.05)
h(19-dst_host_count)
h(dst_host_count-19)
h(3-dst_host_srv_count)
h(dst_host_srv_count-3)
h(0.03-dst_host_diff_srv_rate)
h(dst_host_diff_srv_rate-0.03)
h(dst_host_same_src_port_rate-0.18)
h(0.99-dst_host_same_src_port_rate)
h(dst_host_same_src_port_rate-0.99)
h(0.15-dst_host_srv_diff_host_rate)
h(dst_host_srv_diff_host_rate-0.15)
h(0.16-dst_host_serror_rate)
h(dst_host_serror_rate-0.16)
h(0.98-dst_host_rerror_rate)
h(dst_host_rerror_rate-0.98)
h(0.99-dst_host_srv_rerror_rate)
h(dst_host_srv_rerror_rate-0.99)
```




```
> var_keep
 [1] "dst_host_serror_rate"
 [2] "root_shell"
 [3] "dst_host_diff_srv_rate"
 [4] "serviceeco_i"
 [5] "dst_host_srv_rerror_rate"
 [6] "hot"
 [7] "serviceecr_i"
 [8] "dst_host_same_src_port_rate"
 [9] "su_attempted"
[10] "dst_host_rerror_rate"
[11] "serviceftp_data"
[12] "wrong_fragment"
[13] "dst_host_count"
[14] "num_file_creations"
[15] "diff_srv_rate"
[16] "flagRSTR"
[17] "duration"
[18] "num_failed_logins"
[19] "dst_host_srv_diff_host_rate"
[20] "dst_host_srv_count"
```

MARS Based Ensemble Classification Algorithm

An ensemble algorithm adopts MARS as base classifiers using both Bagging and Random Subspace Method (RSM)

A number of parameters are required to be specified for MARS-based ensemble

- The number of MARS base classifiers to be included in the ensemble
- The number of variables to be selected as random feature subspaces
- The maximum number of basis functions
- The maximum degree of interactions to be used in the smoothing spline estimation

Future Research
Optimization of Ensemble Based Decision Using PSO

- Both experimental and theoretical studies have proved that classifier fusion can be effective in improving overall classification performance.
- Each MRAS produces its predicted class, then the final predictor is obtained by weighted majority vote.
- Particle Swarm Optimization (PSO) technique can be applied to determine the appropriate weights.
- The weights associated to each base classifier on the basis of its accuracy are optimized using the basic idea of PSO.

Future Research
Optimization of Ensemble Based Decision Using PSO

- PSO is a population based stochastic optimization technique which is inspired by social behaviour of birds.
- The particles are flown through the multi-dimensional search space with each particle representing a potential solution to the multidimensional problem.
- Each solution's fitness is based on the objective function related to the optimization problem being solved. To achieve better performance of ensemble classifiers, the fitness function defined as classification error (accuracy) is minimized (maximized).
- In each iteration a new population in the PSO algorithm is updated by shifting the positions of the previous one. During its movement, each particle is influenced by its peer's and its own trajectory.

Reference: Particle swarm optimization algorithm: an overview, Soft Computing · January 2017

Future Research
Optimization of Ensemble Based Decision Using PSO

- PSO makes few or no assumptions about the problem being optimized and can search relative large spaces of potential solutions
- PSO does not require the optimization problem differentiable as required by classic optimization methods such as quasi-newton methods and gradient descent
- PSO is easy to understand, easy to implement, and converges fast
- Improved PSO algorithms avoids premature convergence by offering more control of the parameters

Reference: Particle swarm optimization algorithm: an overview, Soft Computing · January 2017

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Application of Explainable Machine Learning Systems for Improving Network Intrusion Detection

Thank You!

Ying Zhou
George Washington University, Ph.D. Student
ycindyzhou@gwu.edu