Interpreting Unsupervised Anomaly Detection in Security via Rule Extraction

Ruoyu Li, Qing Li, Yu Zhang, Dan Zhao, Yong Jiang, Yong Yang

Paper link

Homepage link

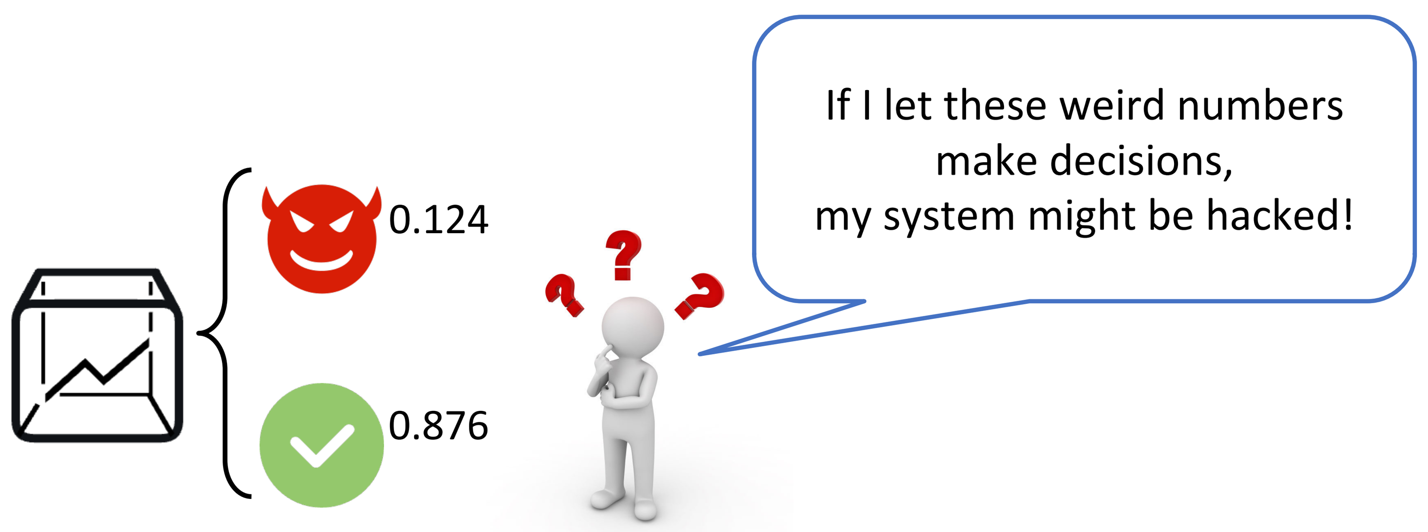
Background

ML/DL Unsupervised Anomaly Detection for Security Applications

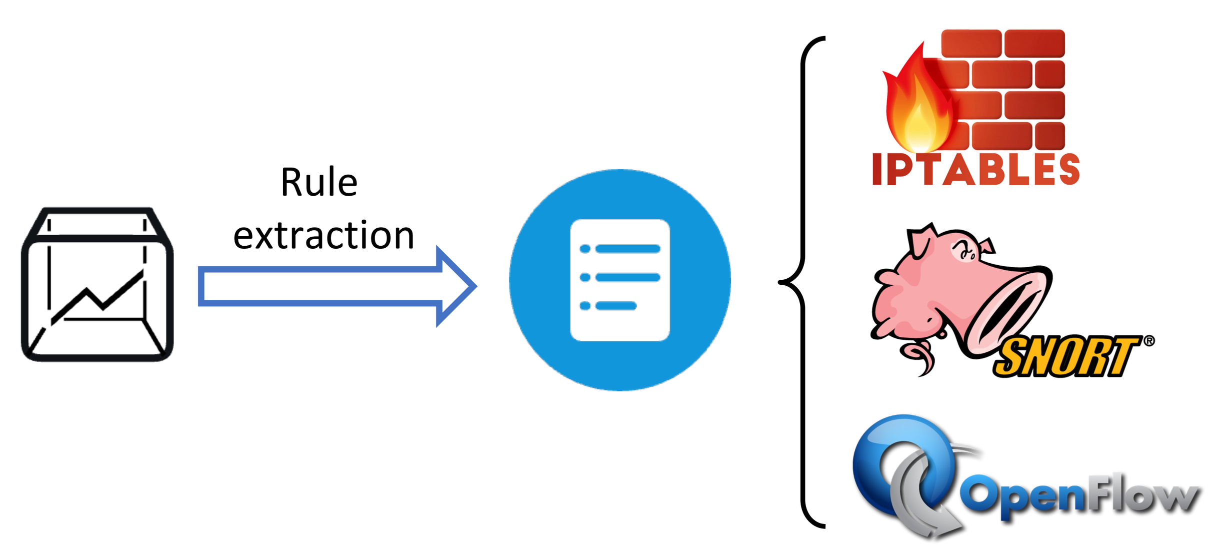
- ✓ VAE, iForest, OCSVM, … ⇒ Network Intrusion Detection, Malware Detection, …
- ✓ Do not require labeled attack data which are extremely sparse
- ✓ Better detection of unforeseen anomalies (e.g., 0-day attack)

Motivation for Rule Extraction

- Trust over High-Stake Security Decisions



- Integration with Online Defense



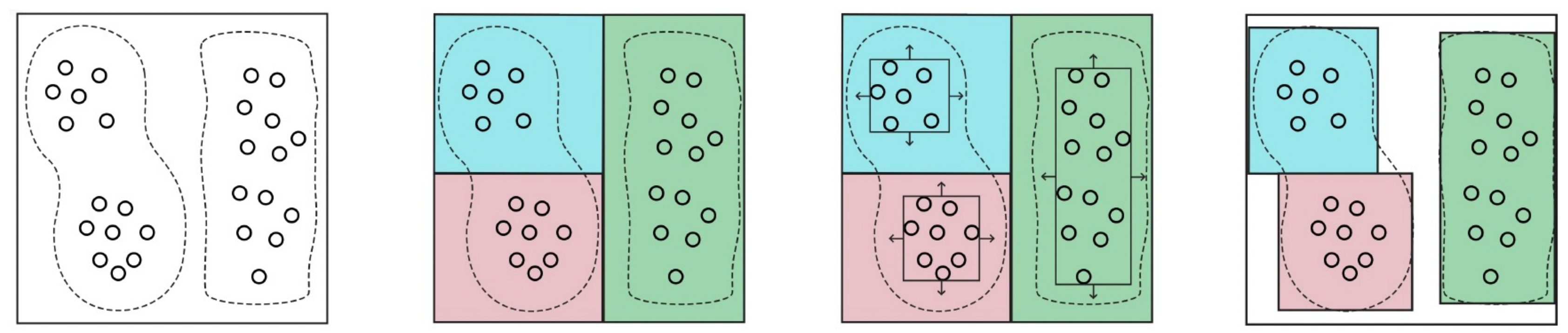
Rule-based defense tools

Why is Globally Explaining Unsupervised Anomaly Detection Challenging?

- **Unlabeled One-class Data**: No labeled data to determine decision boundaries
- **Lack of Surrogate Models**: Most self-explained surrogates (e.g., CART) are supervised
- **Accuracy Loss**: Security applications require high detection accuracy for online defense

Intuition

Normal network data are typically **multimodal**, e.g., a server supports multiple services as normal behaviors, such as web, email and database



(a) The unlabeled data (b) Compositional distributions (c) Process of the CBE algorithm (d) The final rule set

Method: Divide-and-Conquer!

We extract two types of rules:

- **Distribution Decomposition Rule** (Figure b)
- ➢ Obtained by **IC-Tree**: we let *anomaly scores* output by the original detection model guide node splitting

$$p = \mathbb{E}_{x \in N}[P_{\mathcal{X}\sim\mathcal{D}}(\boldsymbol{x})] = \frac{1}{|N|}\sum_{x \in N} f(\boldsymbol{x})$$

$$C_k^I = (x_i \odot_1 b_i | s_1 = (i, b_i)) \wedge \cdots \wedge (x_j \odot_\tau b_j | s_\tau = (j, b_j))$$

- **Boundary Inference Rule** (Figure c)
- ➢ Obtained by **CBE algorithm**: starting from *a* compact *hypercube* that encompasses normal data and *exploring decision boundaries* by a gradient approximation approach

Algorithm 1: Compositional Boundary Exploration

Input: Data falling into the k-th leaf node $\boldsymbol{X}_k$, anomaly detector f and its threshold φ
Output: Boundary inference rule C_k on this leaf node such that C_k encapsulates normality

$H_k \leftarrow \text{MinimalHypercube}(X_k)$;
for i-th dimension **in** $\boldsymbol{X}_k$ **do**
  $\boldsymbol{e}^{(1)}, ..., \boldsymbol{e}^{(N_e)} \leftarrow \text{IntialExplorer}(H_k)$ on i-th dimension;
  while True **do**
    $\hat{\boldsymbol{e}}^{(1)}, ..., \hat{\boldsymbol{e}}^{(N_s)} \leftarrow \text{AuxiliaryExplorer}(\boldsymbol{e})$ for each initial explorer $\boldsymbol{e}$;
    Beam Search for N_e candidate explorers from $N_e \times N_s$ auxiliary explorers that have the minimal probability of being normal judged by f and φ;
    $\boldsymbol{e} \leftarrow \text{GradientApprox}(\hat{\boldsymbol{e}})$ for each candidate explorer selected from auxiliary explorers;
    if ending condition satisfied **then**
      $c_i \leftarrow (x_i \odot \hat{e}_i)$ and **break**;
  end while
end for
return $C_k^E = H_k \vee (c_1 \wedge c_2 \wedge ... \wedge c_d)$;

- Final Rule: the conjunction of two types of extracted rules for each compositional distribution (Figure d)

Evaluation

Quality of Rule Extraction

- Highest **fidelity** on all the detection models and datasets
- Highest **TPR** on all the detection models and datasets
- A high level of **robustness** and **TNR**

CIC-IDS2017 dataset

Method	AE				VAE				OCSVM				iForest			
	FD	RB	TPR	TNR	FD	RB	TPR	TNR	FD	RB	TPR	TNR	FD	RB	TPR	TNR
UAD	0.1325	0.4991	0.0003	0.9792	0.1438	0.4839	0.022	**0.9988**	0.0725	0.5000	0.00	1.00	0.1262	0.5000	0.0	**1.00**
EGDT	0.533	**1.00**	0.4354	0.9947	0.1437	**1.00**	0.022	0.9961	0.9189	0.9994	0.9306	0.838	0.9729	0.9996	0.9417	0.9189
Trustee	0.4871	0.6412	0.3844	0.9981	0.1552	0.9857	0.0152	**0.9988**	0.539	0.6108	**1.00**	**1.00**	0.4543	0.5801	0.9795	0.4486
LIME	0.6918	0.9999	0.7889	0.0014	0.8232	**1.00**	0.9329	0.001	0.068	0.9999	0.0777	0.0241	0.8910	**0.9998**	0.8246	0.9913
KD	0.5776	0.9989	0.4792	**0.9998**	0.2010	0.9817	0.1016	0.9993	0.3620	**1.00**	0.3102	0.9995	0.1262	0.7016	0.00	**1.00**
Ours	**0.9835**	**1.00**	**0.9457**	0.9915	**0.9620**	0.9993	**0.9610**	0.9944	**0.9275**	**1.00**	**1.00**	**1.00**	**1.00**	0.9949	**0.9968**	0.9843

CSE-CIC-IDS2018 dataset

Method	AE				VAE				OCSVM				iForest			
	FD	RB	TPR	TNR	FD	RB	TPR	TNR	FD	RB	TPR	TNR	FD	RB	TPR	TNR
UAD	0.3796	0.3077	0.0004	0.7418	0.2697	0.2930	0.1490	0.4857	0.6051	0.3069	0.3004	0.9876	0.6811	0.4035	0.3539	0.9724
EGDT	0.5821	**1.00**	0.1432	0.9801	0.2197	0.9989	0.2308	0.9554	0.5106	**1.00**	**1.00**	0.9546	0.9546	0.7813	0.9888	0.8971
Trustee	0.5157	0.9006	0.1901	0.9857	0.3642	0.9752	0.0124	0.9636	0.3616	0.5955	**1.00**	**1.00**	0.4241	0.4700	0.9641	0.5162
LIME	0.5838	0.9997	0.7681	0.0255	0.6814	**1.00**	0.9402	0.0213	0.0560	**1.00**	0.9999	0.0186	0.8903	**1.00**	0.9884	0.8745
KD	0.5074	0.9999	0.3562	**0.9979**	0.4234	0.9989	0.1086	**0.9925**	0.3180	0.9967	0.4308	0.1510	0.3596	0.6834	0.0000	**1.00**
Ours	**0.9954**	0.9997	**0.9998**	0.9774	**0.8962**	0.9985	**0.9997**	0.8268	**0.9929**	0.9997	0.9983	0.9753	**0.9947**	0.9291	**0.9988**	0.9583

TON-IoT dataset

Method	AE				VAE				OCSVM				iForest			
	FD	RB	TPR	TNR	FD	RB	TPR	TNR	FD	RB	TPR	TNR	FD	RB	TPR	TNR
UAD	0.1499	0.015	0.0258	0.908	0.2157	0.4010	0.1863	0.7787	0.0489	0.5000	0.00	**1.00**	0.0674	0.5000	0.00	**1.00**
EGDT	0.9750	**1.00**	0.9739	0.9943	0.7660	**1.00**	0.7538	0.9948	0.8139	0.9997	0.8051	0.9759	0.6345	0.9226	0.6247	0.9475
Trustee	0.4774	0.5722	0.4502	0.9971	0.3807	0.6689	0.3484	0.9975	0.7942	0.8430	**1.00**	**1.00**	0.7476	0.8145	0.9824	0.1943
LIME	0.6971	0.9999	0.7939	0.0027	0.8289	**1.00**	0.9379	0.0015	0.0687	0.9999	0.0787	0.0231	0.8963	**0.9998**	0.8296	0.9918
KD	0.0821	**1.00**	0.0341	**0.9987**	0.0591	0.9997	0.0099	**0.9980**	0.0494	**1.00**	0.0005	0.9994	0.0674	0.9955	0.00	**1.00**
Ours	**0.9996**	**1.00**	**1.00**	0.9845	**0.9995**	**1.00**	**1.00**	0.9831	**0.9511**	**1.00**	**1.00**	0.9881	**1.00**	0.9890	**1.00**	0.9715

Understanding Model Decisions

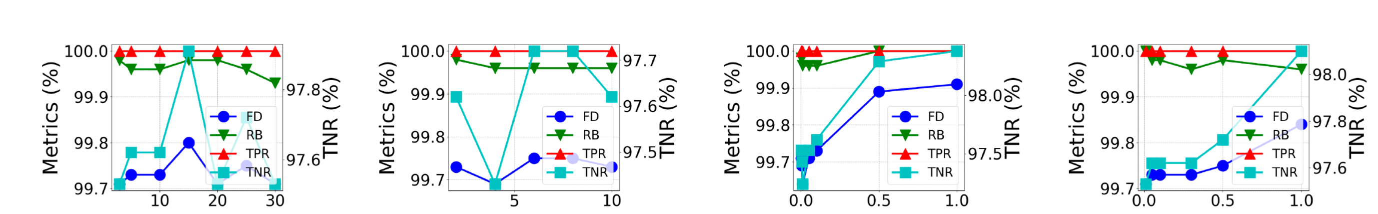
- Explaining how models detect 4 attack types by rules
- Feature values markedly **higher** or **lower** than the bounds of rules
- Explanations are in line with **how humans recognize** the attack data

Attack	Rules of Normality	Attack Value	Feature Meaning	Human Understanding
DDoS	ps_mean $>$ 101.68 iat_mean $>$ 0.063 dur $>$ 12.61	57.33 0.00063 0.00126	Mean of IP packet sizes Mean of packet inter-arrival time Duration of a connection	DDoS attacks use packets of small sizes to achieve asymmetric resource consumption on the victim side, and send packets at a high rate to flood the victim.
Scanning	count $>$ 120 ps_var $>$ 2355.20	1 0.0	IP packet count per connection Variance of IP packet sizes	Scanning attacks send a constant probe packet to a port, and the victim will not reply if the port is closed.
SQL Injection	ps_bwd_mean $\leq$ 415.58 dur $>$ 1.64	435.80 0.37	Mean of backward IP packet sizes Duration of a connection	Unauthorized access to additional data from websites, usually establish short connections for one attack.
Backdoor	ps_max $>$ 275.28 ps_min $>$ 49.41	48.0 40.0	Maximum of IP packet sizes Minimum of IP packet sizes	It persists in compromised hosts and sends stealthy keep-alive packets with no payload (thus very small).

Computational Complexity and Hyperparameter Sensitivity

Table 6: Average training and prediction time per sample for different feature sizes.

Feature Size	Training Time (ms)	Prediction Time (ms)
20	$5.40 \pm 5.50 \times 10^{-4}$	$5.48 \times 10^{-3} \pm 2.51 \times 10^{-9}$
40	$15.5 \pm 6.80 \times 10^{-2}$	$5.52 \times 10^{-3} \pm 2.34 \times 10^{-9}$
60	$14.7 \pm 8.75 \times 10^{-5}$	$6.99 \times 10^{-3} \pm 3.56 \times 10^{-8}$
80	$30.7 \pm 3.08 \times 10^{-1}$	$6.91 \times 10^{-3} \pm 9.00 \times 10^{-8}$



(a) Maximum tree depth (b) #Explorers (c) Sampling coefficient (d) Iteration stride