Classification of Malicious Cyber Activity Via Machine Learning Against Event Logs

SAND2021-10489PE

Authors: Joseph Cutchall (NMT B.S. Candidate, CSE), Wesley Pick-Roth (NMT M.S. Candidate, CSE), Max Planck (NMT PhD Candidate, CSE; ICASA Technical Director), Nathanael Brown (Sandia, 8721)

NM Partnerships

LDRD U.S. DEPARTMENT OF ENERGY NNSA

Fall 2021 LDRD Virtual Poster Session

Introduction and Motivation

- NMT is exploring novel mechanisms for detecting malicious activity in system log traffic.
- The project is an extension to an initial collaboration with SNL in which basic research and preliminary experiments were conducted.
- **Overall Motivation:** Look for mechanisms for detecting/classifying non-obvious malicious activity in systems by leveraging existing and/or emerging machine learning (ML) techniques.
  - *Multi-faceted, team approach*: several NMT students and researchers are working on distinct, yet highly related approaches
  - *Big Picture:* work towards the application of ordered-policy memory networks
- **Guiding Principles:**
  - Work with real-world or at least emulated data sets
  - Continuously improve data pre-processing and refinement while successively working towards more complex ML regimes (in particular, Ordered Memory Policy Networks, or OMPNs)

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND Release SAND 21-XXXX

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Classification of Malicious Cyber Activity Via Machine Learning Against Event Logs

Authors: Joseph Cutchall (NMT B.S. Candidate, CSE)), Max Planck (NMT PhD Candidate, CSE; ICASA Technical Director), Nathanael Brown (Sandia, 8721)

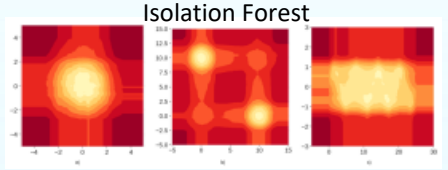
NM Partnerships

Fall 2021 LDRD Virtual Poster Session

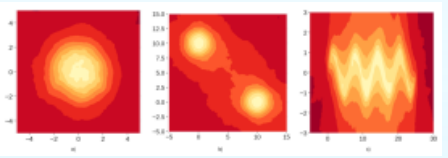
Approach

- **Extended Isolation Forests (EIFs)**
  - EIFs are an addition to the IF algorithm that is made to tackle the problem Isolation forests had with contamination of scoring data, that can lead to misclassification of outliers.
- **Approach**
  - Previous work: Use of Isolation Forests (IFs) against log data
  - Learning: validate previous IF experiment results
  - Development: expand code base to accommodate EIFs
  - Application: perform previous experiments with EIFs
  - Analysis: compare results / iterate over parameters as necessary
- **Data**
  - ICASA internal data, produced from its Process Analysis for Cyber Operations project (PACO)
  - (Recently available) TracerFIRE 9 (TF9) data
- **Supplemental**
  - Refinement of existing data (e.g., feature space)
  - Pre-processing of TF9 data

Isolation Forest



Extended Isolation Forest



Pictures From https://github.com/sahandha/eif#Use

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.. Unlimited Release SAND 21-XXXX

Classification of Malicious Cyber Activity Via Machine Learning Against Event Logs

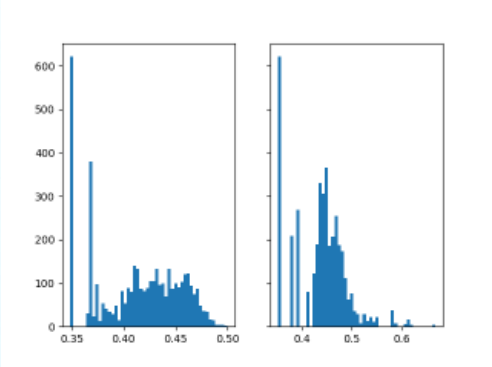
Authors: Joseph Cutchall (NMT B.S. Candidate, CSE), Max Planck (NMT PhD Candidate, CSE; ICASA Technical Director), Nathanael Brown (Sandia, 8721)

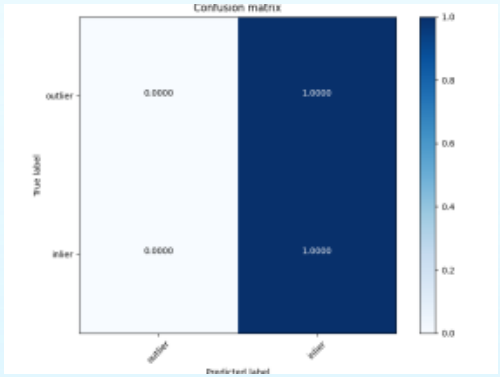
NM Partnerships

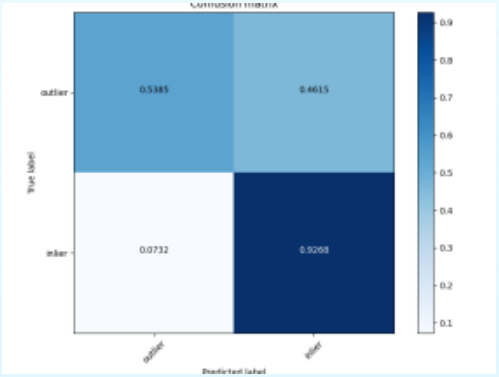
Fall 2021 LDRD Virtual Poster Session

Current Status and Results (if any)

- EIFs against PACO data
  - Initial Results Show that the EIFs were able to better detect outliers while generally reducing the number of false positives reported.
  - With one specific run we were able to see a 53.8% increase in event detection while only increasing the false positive reporting of 6.5%
- TF9 Data
  - We are currently working on preprocessing TF9 Data to be able to run it through IFs and EIFs







This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.. Unlimited Release SAND 21-XXXX

NM STATE All About Discovery!™ New Mexico State University

THE UNIVERSITY OF NEW MEXICO

NEW MEXICO TECH SCIENCE · ENGINEERING · RESEARCH · UNIVERSITY

Authors: Joseph Cutchall (NMT B.S. Candidate, CSE), Max Planck (NMT PhD Candidate, CSE; ICASA Technical Director), Nathanael Brown (Sandia, 8721)

Impact of Work

- **Conclusions:**
  - With using Extended Isolation Forests on the Paco Data set we were able to get better results in classifying bad actors from the data set. Though we were not able to classify all of the bad actors in the data. With the Implementation of EIFs it seems that to further our results we need to find a way for bad actor data to be better represented as anomalous after preprocessing.

- **Impact of work:**
  - Comprehensive analysis of methodology against available data (building on previous cursory experiments)
  - Parameter sweeps are helping drive better understanding of data
  - In turn, will help to inform future experiments

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.. Unlimited Release SAND 21-XXXX

Classification of Malicious Cyber Activity Via Machine Learning Against Event Logs

Authors: Wesley Pick-Roth (NMT M.S. Candidate, CSE), Max Planck (NMT PhD Candidate, CSE; ICASA Technical Director), Nathanael Brown (Sandia, 8721)

NM Partnerships

Fall 2021 LDRD Virtual Poster Session

Approach

- **Isolation Forests**
  - Determine outliers by randomly partitioning data
  - Likelihood of outlier proportional to ease of isolation
- **Approach**
  - Previous work: Use of Isolation Forests (IFs) against log data
  - Learning: Validate previous IF experiment results
  - Development: Explore different encoding strategies of datasets
    - Incorporation of timestamp in encoded events
    - Alternative encodings of *image* (executable) field
  - Application: Perform previous experiments with IFs using alternative encodings
  - Analysis: Compare results / iterate over parameters as necessary
- **Data**
  - ICASA internal data, produced from its Process Analysis for Cyber Operations project (PACO)
  - (Recently available) TracerFIRE 9 (TF9) data
- **Supplemental**
  - Pre-processing of TF9 data

Isolation of a normal point

Isolation of an anomaly

Image of isolation forest partitioning from https://laptrinhx.com/isolation-forest-algorithm-for-anomaly-detection-3651893833/

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.. Unlimited Release SAND 21-XXXX

Classification of Malicious Cyber Activity Via Machine Learning Against Event Logs

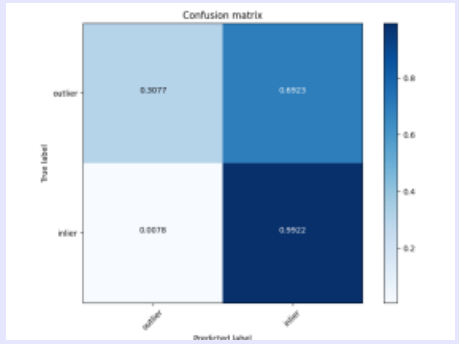
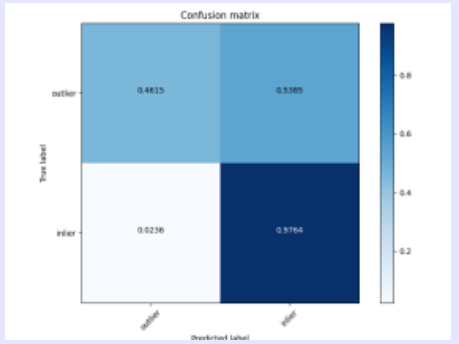
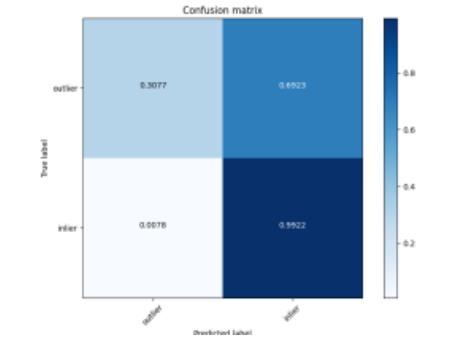
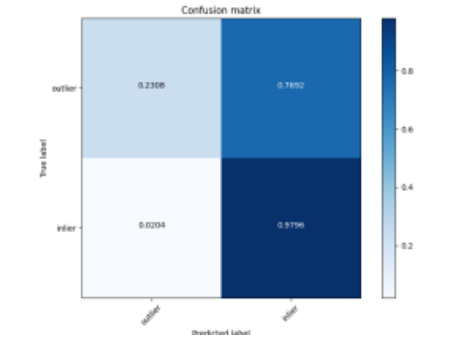
Authors: Wesley Pick-Roth (NMT M.S. Candidate, CSE), Max Planck (NMT PhD Candidate, CSE; ICASA Technical Director), Nathanael Brown (Sandia, 8721)

NM Partnerships

Fall 2021 LDRD Virtual Poster Session

Current Status and Results (if any)

- **Incorporation of Timestamps**
  - Include ‘time since epoch’ and/or ‘time since midnight’ for time relation
  - Finds more anomalous data at expense of more misclassed safe data
  - Image explanation:
    - True positive rate - 0.3077 to 0.4615
    - True negative rate - 0.9922 to 0.9764
- **Change of *image* encoding**
  - Change *image* data from single numeric feature to one-hot encoding
  - Contrary to earlier work, performs better with smaller parameterization
  - Image explanation:
    - True positive rate – 0.3077 to 0.2308
    - True negative rate – 0.9922 to 0.9736
  - Not tuned

Original	Modified
	
	

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.. Unlimited Release SAND 21-XXXX

Authors: Wesley Pick-Roth (NMT M.S. Candidate, CSE), Max Planck (NMT PhD Candidate, CSE; ICASA Technical Director), Nathanael Brown (Sandia, 8721)

Impact of Work

- **Conclusions:**
  - Implementation of an order of events needs further exploration
  - New image encoding gives promising results without tuning

- **Impact of work:**
  - Reevaluating data features may lead to further improvements
  - Fewer learning parameters adds more generality
  - Evaluating parameters provided insight to future work
    - Use sequence of process events, rather than time
  - In turn, will help to inform future experiments (as detailed in the next slide)

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.. Unlimited Release SAND 21-XXXX

Classification of Malicious Cyber Activity Via Machine Learning Against Event Logs

Authors: Joseph Cutchall (NMT B.S. Candidate, CSE), Wesley Pick-Roth (NMT M.S. Candidate, CSE), Max Planck (NMT PhD Candidate, CSE; ICASA Technical Director), Nathanael Brown (Sandia, 8721)

NM Partnerships

Fall 2021 LDRD Virtual Poster Session

Challenges and Risks / Next Steps and Future Work

- Challenges / Risks
  - EIFs seem to run much less efficiently than regular IFs
  - Parameter tuning needed for best results
  - Scaling may be an issue for both paradigms presented above
  - Heavy reliance on preprocessing of Data

- Next Steps:
  - Run TF9 data through IFs and EIFs
  - Build graphs from event relations in data and run metrics as features in the Forests
  - Extract graph metrics / fuse with existing ‘intrinsic’ feature set and re-run previous experiments

- Future Work:
  - Use graph data to construct event ‘sequences’ and develop labeling
  - Use a sequence-based learning system (e.g. LSTM, HMM) as a building block towards OMPNs
  - Experiment with OMPNs in an attempt to abstract sequence data to labels that align with malicious activity (e.g., the cyber-kill chain)

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.. Unlimited Release SAND 21-XXXX