Summary Analytics

Cybersecurity

Summarization and prioritization of cybersecurity records

See it online!

www.smr.ai

Problem statement

- Worldwide spending on information security products and services exceeded **$114 billion in 2018**, an increase of 12.4 percent from 2017, according to Gartner, Inc. For 2019, they forecast the market to grow to $124 billion, and **$170.4 billion in 2022**.[1]
- Yet, according to the Dimensional Research[2]:
  - **96%** of companies have run into training related problems – including data quality, labeling required to train an AI system, and building model.
  - **78%** of AI/ML projects stall at some stage before deployment.
- Processing the evergrowing torrents of security event data is extremely costly:
  - For Machines:
    - Processing, training, and storage of data has appreciable financial and environmental costs.
  - For Humans:
    - Alert and decision fatigue due to data abundance and redundancy.

[1] https://cybersecurityventures.com/cybersecurity-market-report/
[2] https://content.alegion.com/dimensional-researchs-survey

Solution and product

- While cybersecurity data is large, it is also **redundant** and **imbalanced,** and this can be exploited to save time, reduce costs, and improve information extraction.

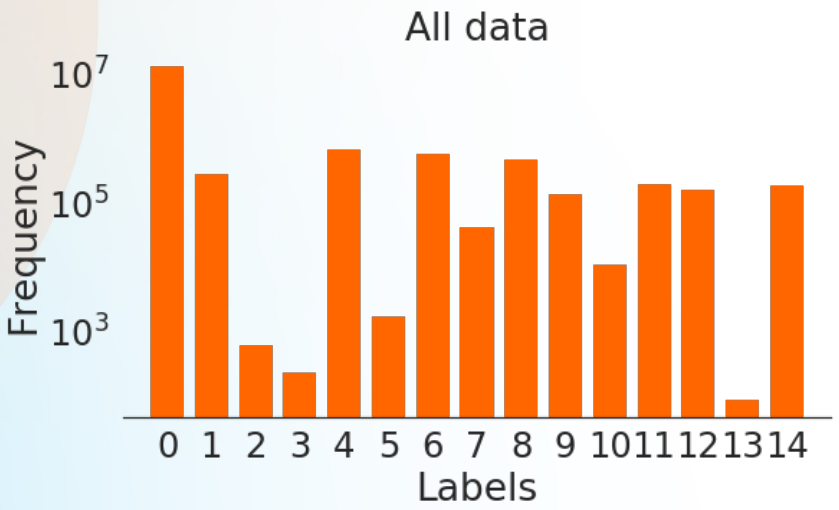
Solutions: SMRaiz and LINKaiz

- **SMRaiz eliminates redundancy and restores balance to the data**. This therefore:
  - Reduces AI/ML training costs;
  - Reduces data labeling costs and errors;
  - Reduces human analyst fatigue and errors and improve productivity; and
  - Reduces data bias and imbalance.

- **LINKaiz links redundant alerts to those prioritized in the summary.** This therefore:
  - Identifies all instances of security threats quickly and cost efficiently; and
  - Reduces false positives and false negatives efficiently.

Example

- Task
  - Summarization of massive collection of Intrusion Detection System (IDS) records
- Dataset
  - CSE-CIC-IDS2018: https://www.unb.ca/cic/datasets/ids-2018.html
  - Consists of labeled network flows (IDS records) covering 15 different categories: 1 benign traffic and 14 different types of attack scenarios (see the last slide for the exact labels)
  - Featurized using standard featurization methods (CICFlowMeter-V3)
  - Dataset size = 16,137,183 records
  - Extremely imbalanced traffic type:
    - ~83% benign (label 0)
    - ~17% different attack types (label 1– 14)



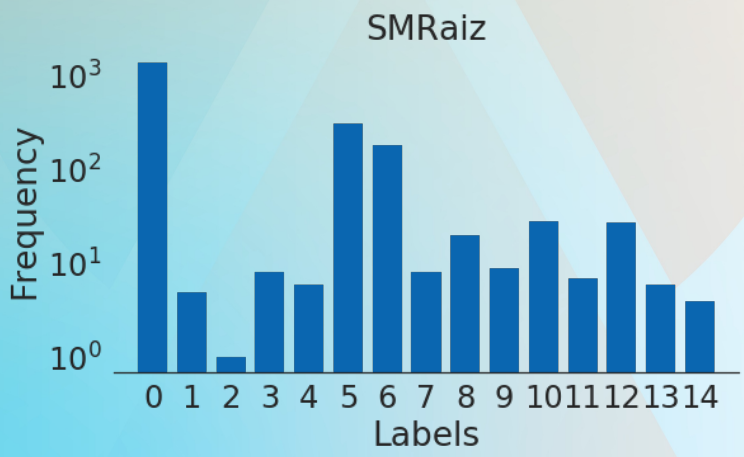
See it online!

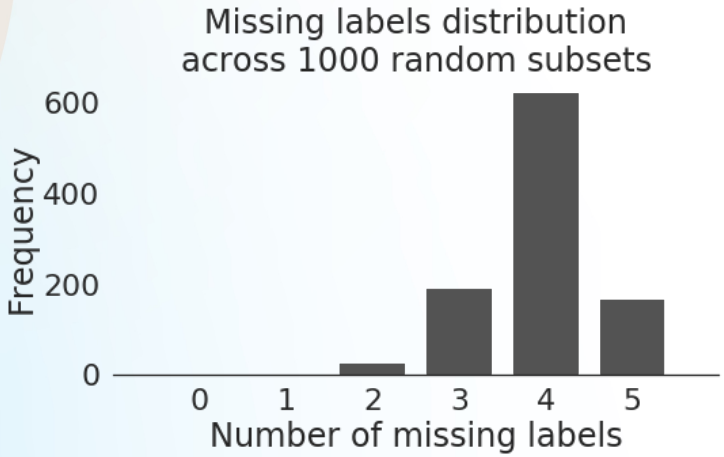
Cybersecurity record types

- **IDS records types:**
  - Benign – 83% of the data.
  - Fourteen attack types – 17% of the data:
    1. BotNet
    2. Brute Force –Web
    3. Brute Force –XSS
    4. DDOS attack-HOIC
    5. DDOS attack-LOIC-UDP
    6. DDoS attacks-LOIC-HTTP
    7. DoS attacks-GoldenEye
    8. DoS attacks-Hulk
    9. DoS attacks-SlowHTTPTest
    10. DoS attacks-Slowloris
    11. FTP-BruteForce
    12. Infiltration
    13. SQL Injection
    14. SSH-Bruteforce

Results

- **Results**
  - Data is summarized down to 0.01% (2000 elements, more than a 10,000x reduction!). Summarization does not know attack types beforehand.
  - The label distribution is reported for the summary set. Also 1000 unbiased random subsets (each of size 2000 elements) were produced.
  - All 1000 random subsets miss **at least 2 attack types**, with the majority missing **4 or more attack types**! Unbiased random selection fails!
  - SMR.ai's summary misses **nothing**! Captures all 15 categories!
  - Summarization time = 66.03 sec (single thread, Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz)





[See it online!]

Implications and benefits

- **Human data analysts**
  - Analyzing the summary set, which is representative of the original, is much easier, faster, more cost-effective.
  - Unbiased random subsets would miss important attack types!
- **Efficient AI model development**
  - AI model search and hyperparameter tuning are costly and time consuming.
  - Training AI models on summary sets allows for more cost-effective AutoML and Network Architecture Search (NAS) techniques.
- **Bias and imbalance reduction**
  - A summary set (unlike a random subsample) is necessarily diverse, and free from inherent bias present in the entirety.
  - A summary de-biases and rebalances the data relative to the entirety.
  - With a summary, one can obtain models with accuracies converging to the entire set and potentially exceeding it because of a better class balance and reduced bias.