Prediction-Based Cyber Analytic Threat Detection

Kedrick Hill

The Big Picture

- 80% of businesses were successfully hacked in 2015 [CBS MoneyWatch]
- Detecting misuse of data efficiently and accurately
- One leading development is through machine learning with prediction algorithms

Outline

1. Background
    a. Machine Learning
    b. Predictive Analytics
    c. Cybersecurity
2. Clustering
3. Decision Trees
4. Support Vector Machines
5. Hybridization
6. Conclusion

Background: Machine Learning

- The process of a system to learn through experience
- Uses Data Science and Data Mining techniques
- Two most common Learning Types:
  - Supervised Learning:
    - Algorithm learns through an outcome
    - Uses labeled (tagged w/ classifications) training data
      - Used to fit model
  - Unsupervised Learning:
    - Analyzes data & learns patterns w/o outcome
    - Data is unlabeled

Background: Predictive Analytics

- A form of business analytics that predicts an outcome through data
- Predictive models (graphical or non-graphical) can calculate patterns or trends from past, current, or future data
- Use of pattern recognition detects anomalies when monitoring or detecting attacks

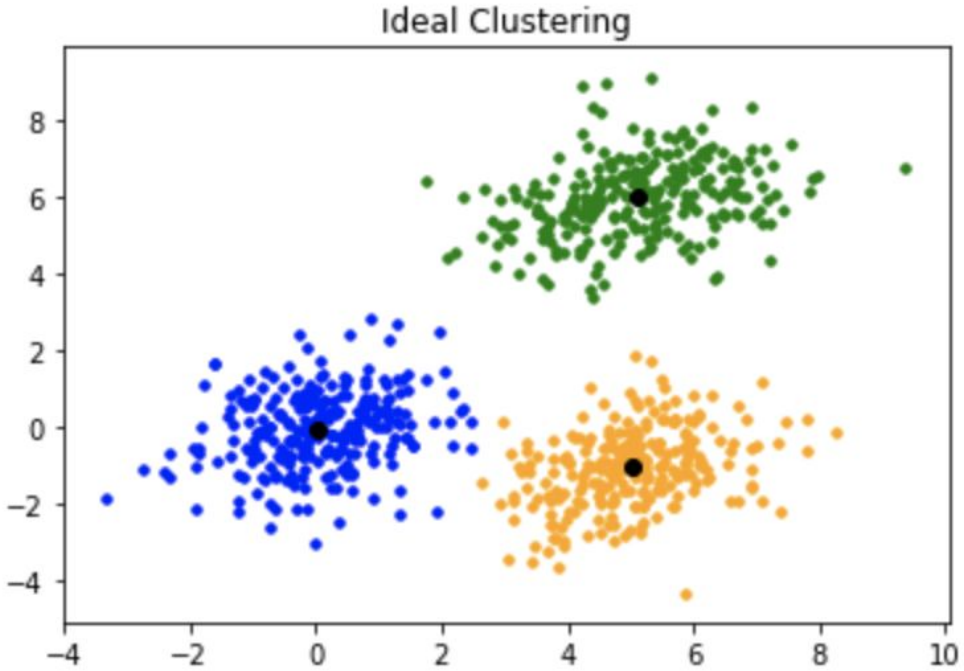
Background: Cybersecurity

- The process of ensuring information system protection including software, hardware, and any information or data
- Businesses struggle the most with:
  - Attacks
  - Cyber espionage
  - Data theft
- Cyber security intends to achieve:
  - Data confidentiality
  - Availability
  - Stronger Authentication and Integrity

Clustering and K-Means

Clustering

- Pattern recognition method
- Unsupervised
- Buckets data based on their Euclidean distance, or magnitude
- Main advantages in intrusion detection is:
  - Learns from audit (IT infrastructure) data
  - Does not need explicit descriptions
  - Classifies attacks from the data
- Two equations that are commonly used to cluster data:
  - K-Means
  - KNearestneighbor



- Centroids: average position

Example of Clustering

Clustering: K-Means

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - \mu_j\|^2$$

- J - all sets of points
- k - spatial clusters, # of clusters
- n - # of observations
- j - current set counter
- i - current observation value in a set
- x - observation, denoted by a i value
- μ - mean of points set j

J = {(0,7), (6,8), (7,8), (1,5), (1,3), (0,5), (5,6), (2,3), (0,5)}

J sets and their values

	x	y
A	0	7
B	6	8
C	7	8
D	1	5
E	1	3
F	0	5
G	5	6
H	2	3
I	0	5

Centroid sets (initial)

	x	y
ABC	4.33	7.67
DEF	0.67	4.33
GHI	2.33	4.67

- X-position refers to the file trying to be accessed
- Y-position refers to the # of times failed
- Centroid sets are calculated as the mean.

$ABC_X = 0 + 7 + 6 = 13/3 = 4.33$
$ABC_Y = 7 + 8 + 8 = 23/3 = 7.67$

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - \mu_j\|^2$$

	A	B	C	D	E	F	G	H	I
Att 1	19.2	2.9	7.24	18.22	32.9	25.88	3.24	27.24	25.92
Att 2	7.58	60.56	53.54	0.56	1.88	0.9	21.54	3.54	0.9
Att 3	10.86	41.88	32.9	1.88	4.56	5.54	8.9	2.9	5.54

Legend
Yellow - Attack 1
Green - Attack 2
Teal - Attack 3

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - \mu_j\|^2$$

	A	B	C	D	E	F	G	H	I
Att 1	19.2	2.9	7.24	18.22	32.9	25.88	3.24	27.24	25.92
Att 2	7.58	60.56	53.54	0.56	1.88	0.9	21.54	3.54	0.9
Att 3	10.86	41.88	32.9	1.88	4.56	5.54	8.9	2.9	5.54

Legend
Yellow - Attack 1
Green - Attack 2
Teal - Attack 3

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - \mu_j\|^2$$

Cluster groups:
1. BCG
2. ADEFI
3. H

	A	B	C	D	E	F	G	H	I
Att 1	19.2	2.9	7.24	18.22	32.9	25.88	3.24	27.24	25.92
Att 2	7.58	60.56	53.54	0.56	1.88	0.9	21.54	3.54	0.9
Att 3	10.86	41.88	32.9	1.88	4.56	5.54	8.9	2.9	5.54

Clustering: Results

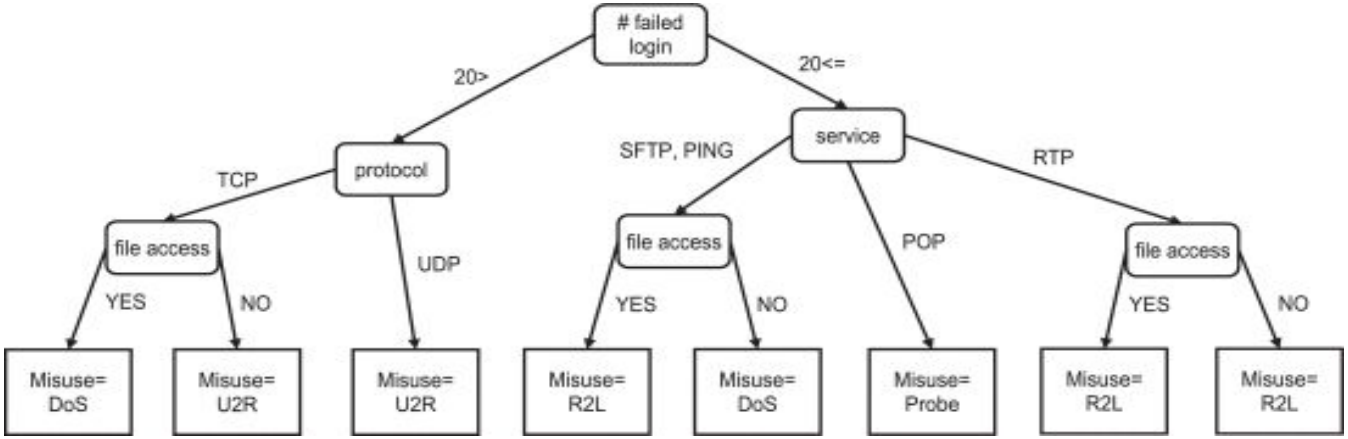
Clustering studies:

- Blowers and Williams did a study on network packets [Intrusion Detection Survey]
  - Partitions the packets into normal or anomalous
  - Performance: 98% accuracy (attack or non-attack)
- Sequeira and Zaki did a study on shell commands at Purdue University [Intrusion Detection Survey]
  - 500 sessions captured
  - Partitioned sessions into regular and intruder
  - Max sequence length: 20
  - Performance: 80% accuracy with 15% false acceptance rate (false positive)

Decision Trees

Decision Trees

- Tree like structures
- Contain attribute or classification nodes
- Uses an input and output method
- Bottom row of nodes are the final node attributes (contain no children nodes)
- Often prebuilt using a SVM
- Supervised

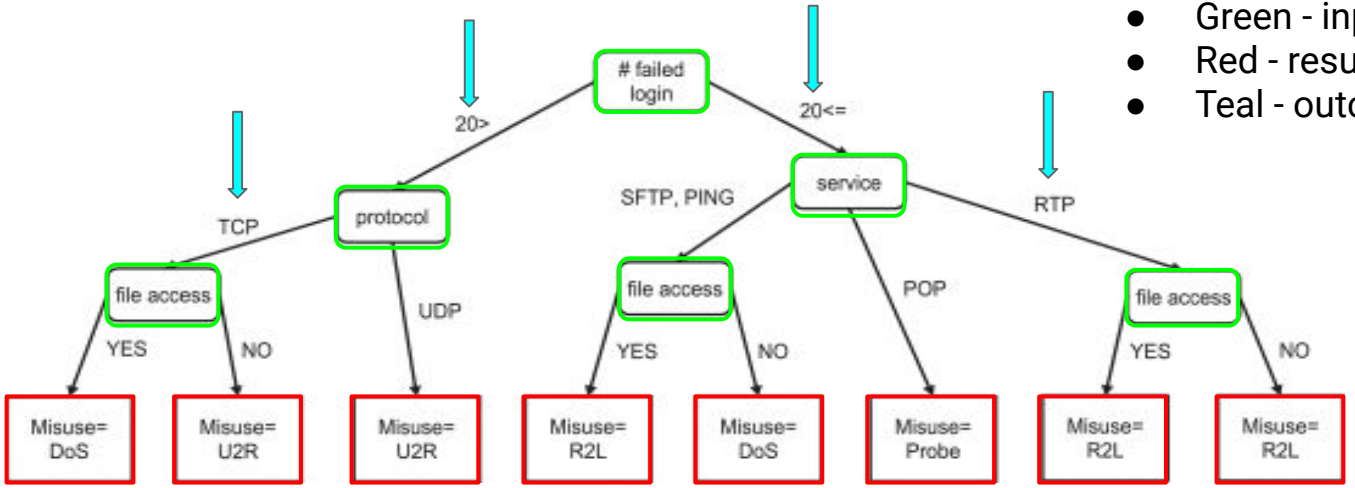


TCP - Transmission Control Protocol
UDP - User Datagram Protocol
SFTP - SSH File Transfer Protocol
PING - Ping Flood DoS
POP - Post Office Protocol
RTP - Real-Time Transport Protocol
DoS - Denial of Service
U2R - User to Root
R2L - Remote to Local

Example of a Decision Tree – Small

Legend:

- Green - input nodes
- Red - result nodes
- Teal - outcome variables



TCP - Transmission Control Protocol
UDP - User Datagram Protocol
SFTP - SSH File Transfer Protocol
PING - Ping Flood DoS
POP - Post Office Protocol

RTP - Real-Time Transport Protocol
DoS - Denial of Service
U2R - User to Root
R2L - Remote to Local

Example of a Decision Tree - Small

Decision Trees: Results

Studies with Decision Trees:

- Kruegel and Toth did a study on Snort, an open source tool [Intrusion Detection Survey]
  - Performed clustering rules to create a tree
  - Studied on tcpdump files from 1999 DARPA evaluation
  - Increasing number of rules increased the speed of the tree
  - Found that clustering methods coupled with decision trees reduce processing time
- Relan and Patil did a study the 2 two KDD data sets (Cup 99 & NSL-) [ML & DL in Cyber]
  - Millions of lines of data in the set
  - Network intrusion dataset
  - Two variations of Decision Trees (w/ and w/o pruning)
    - Prevents overfitting (extra parameters)
  - Found that using pruning had a higher accuracy that w/o using it
  - 98.45% accuracy with a 1.55% false acceptance rate

Support Vector Machines

Support Vector Machines

- Accurate, robust, and reliable machine learning algorithm
- Effective when features are high and data points are low
- SVM’s plot data on a high dimensional space
- Supervised

Support Vector Machines: Results

Studies of in anomaly detection:

- Wagner et al. did a study on NetFlow data [Intrusion Detection Survey]
  - Studied record traffic volume in a window kernel
  - Used internet service provider sources
  - Multiple test reported a range:
    - 89% - 94% accuracy on various attacks
    - 0% - 3% error rate
- Perez and Farid did a study on Network Intrusion Data [ML & DL in Cyber]
  - Used NSL-KDD data set
  - Filtering algorithm
  - Tested various feature sizes (3, 36, and 41)
    - 3: 91% accuracy
    - 36: 99% accuracy
    - 41: 99% accuracy

Hybridization

Hybridization

Using multiple methods can . . .

- Increase processing speeds
  - Clustering + Decision Trees
- Simplify methods
- Aide in creation of models
  - SVM -> Decision Tree
- Cover weaknesses of alternate method(s)

Hybridization: Results

Study with Hybridization:

Yeborah-Ofori and Boachie studied the usage of ML and PA algorithms in threat detection

- Used Logistic Regression(LR), Majority Voting(MV), Support Vector Machines(SVM), and Decision Trees(DT)
- Using LR, MV, and SVM to build a DT
- Alternated tests that coupled one of the algorithms with a DT
- DT had best accuracy at predicting attacks

[Malware attacks]

Conclusion

- Cyber security is growing quickly with the high reliance on technology
- Machine learning and predictive analytics lead a new frontier in cybersecurity
- Studies prove ML and PA methods have high accuracy and speeds
- Combining methods can reduce processing time when detecting attacks
- Hybridization can improve anomaly detection

Questions?

References

Yeboah-Ofori and C. Boachie, "Malware Attack Predictive Analytics in a Cyber Supply Chain Context Using Machine Learning," 2019 International Conference on Cyber Security and Internet of Things (ICSIoT), Accra, Ghana, 2019, pp. 66-73, doi: 10.1109/ICSIoT47925.2019.00019.

Kuan-Ching Li, Beniamino Di Martino, Laurence T. Yang, Qingchen Zhang, "Smart Data: State-of-the-Art Perspectives in Computing and Applications," CRC Press, Mar 19, 2019, pp. 113-132.

Y. Xin *et al.*, "Machine Learning and Deep Learning Methods for Cybersecurity," in *IEEE Access*, vol. 6, pp. 35365-35381, 2018, doi: 10.1109/ACCESS.2018.2836950.

R. Adlakha, S. Sharma, A. Rawat and K. Sharma, "Cyber Security Goal's, Issue's, Categorization & Data Breaches," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, 2019, pp. 397-402, doi: 10.1109/COMITCon.2019.8862245.

Wikipedia contributors. "Machine learning." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 4 Sep. 2020. Web. 8 Sep. 2020.

Wikipedia contributors. "Predictive analytics." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 29 Aug. 2020. Web. 8 Sep. 2020.

H. M. Farooq and N. M. Otaibi, "Optimal Machine Learning Algorithms for Cyber Threat Detection," 2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim), Cambridge, 2018, pp. 32-37, doi: 10.1109/UKSim.2018.00018.

A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, Secondquarter 2016, doi: 10.1109/COMST.2015.2494502.

O. Yavanoglu and M. Aydos, "A review on cyber security datasets for machine learning algorithms," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 2186-2193, doi: 10.1109/BigData.2017.8258167.

Wikipedia contributors. "Data science." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 17 Sep. 2020. Web. 22 Sep. 2020.