Intrusion detection and reinforcement learning with SafeML

*

Sudha Vijayakumar

MSSE, San Jose State University
SAN JOSE, USA
sudha.vijayakumar@sjsu.edu

***Abstract*—This paper is centered around a security-risk classification problem while controlling the action of a machine learning model outcomes when required. A safety monitor will be used to verify the decisions of the classifiers. Safety monitors will alert human agents to review and take appropriate action for potentially high-risk security attacks. The confidence of the predictions computed using distance-based metrics will be used by the safety monitor to decide the next course of action - allow request or alert security agent. The feedback from the human agent will serve as a label for the request ('benign' or 'attack' type).**

***Index Terms*—human labeling, safety machine learning, reinforcement learning, security applications, distance metrics, intrusion detection.**

I. Introduction

Safe Machine learning has an active role to play in the decision making of model outcomes, especially in safety-critical applications like Intrusion Detection Systems. Applications of artificially intelligent systems are spanning across diverse domains at a very rapid pace. With the rapid growth and autonomous actions of AI-systems, there arises a need for controlling the outcomes of these systems.

II. Overview of Research Problems

A. Motivation

Not dealing with the undesirable outcomes of AI Systems can be disastrous and harmful. Undoubtedly, AI Safety is an essential component for safety-critical applications involving human life. So, it is implicit that when AI advances, AI Safety should be improved at a similar pace.

B. Problem description

Just like humans, one of the complex real-time scenarios for an AI-based agent is dealing with situations never experienced before. Any wrong intuitions/ assumptions made by the AI-agent under such circumstances can lead to disastrous outcomes. Fig. 1. outlines the five [1] concrete problems in AI Safety. This paper will focus on solving the'Robustness to distributional shift' problem in network intrusion systems using a [2] SafeML classifier based on different distance metrics.

Fig. 1. Concrete problems in AI Safety

Let's understand this more with an example: Garbage collecting robots could not identify different objects and their significance for the first time. It will assume every item in the trash to be 'trash' even if it is something precious like a diamond. Say, for example, the robot encounter a diamond for the first time, there is no way for it to understand the value of the item. For every new item seen, the robot shall be trained on how to deal with them. This learning is more like humans trying to be proactive to avoid mistakes for unseen situations.

III. Proposed solution

The focus of this paper would be exploring and experimenting with a potential solution to deal with the AI safety issue 'Robustness to distributional shift'. Training the AI-agent to learn new scenarios through [3] reinforcement learning and take safe actions in the future steps will be the goal of this

experiment. For every unseen safety-critical circumstance that the AI-agent encounters, there will be human interference to transfer this active learning to the AI-agent. Experimental code available at https://github.com/sudha-vijayakumar/CMPE257-SafeML-IntrusionDetectionSystem

IV. Technical Details

The SafeML idea proposed by [1], shall be applied to make safe decision making whenever the classifiers come up with a low confidence score for unlabeled data.

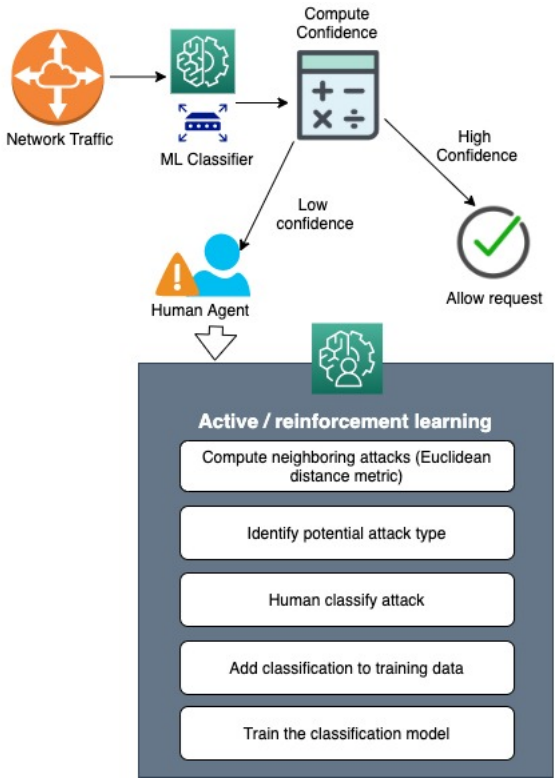

Fig. 2. High-level system design

A. Classify attack

ML Classifier shall classify network data into four different categories - Benign, Attacks(DDoS, Bot, PortScan).

B. Compute prediction confidence

As shown above, the classifier model outcome will be the 'confidence' rate of prediction based upon which further action will be taken. If the confidence level is less than the defined threshold, control is transferred to the human agent to be reviewed for labeling.

C. Find similar attacks and label low-confidence attack

The [4] human agent uses [5] KNN-classification with different distance metrics to find similar attacks and classify the [6] unlabeled data accordingly. Similar attacks are found using distance metrics with the statistical data parameters and data distribution,

Metric	Equation	Training time (secs)	Confidence
Euclidean	$\sum_{i=1}^{n} \sqrt{(Pi - Qi)^2}$	46	85.38
Chebychev	$\max(\lvert x1 - x2 \rvert, \lvert y1 - y2 \rvert)$	58	74.81
Manhattan	$\sum_{i=1}^{n} \lvert Pi - Qi \rvert$	60	85.38
Minkowski	$\sum_{i=1}^{n} \lvert Pi - Qi \rvert ^ p$	59	85.38

Fig. 3. Summary of KNN classification results

V. Experimental Study - Intrusion Detection System

[7] CICIDS2017 Dataset provided by Canadian Institute for Cybersecurity is used to conduct the proposed experiment. There are total of 703245 samples and 79 features. PCA dimensional reduction techniques are applied to extract the features that are contributing most variance towards the attack classification.

A. Types of network data

There are 4 distinct attack types in the used dataset. K-Nearest Neighbour classifier is used to classify the attack types where the optimal number of neighbors is obtained by using the 'elbow curve' method.

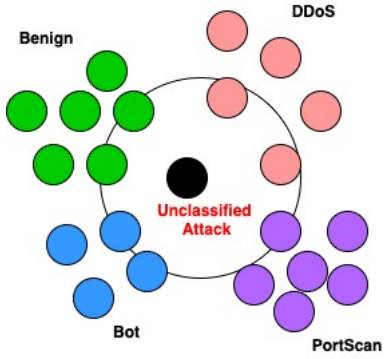

Fig. 4. Types of attack

Type	Size
Benign	414322
Portscan	158930
DDoS	128027
Bot	1966

Fig. 5. Data size and description

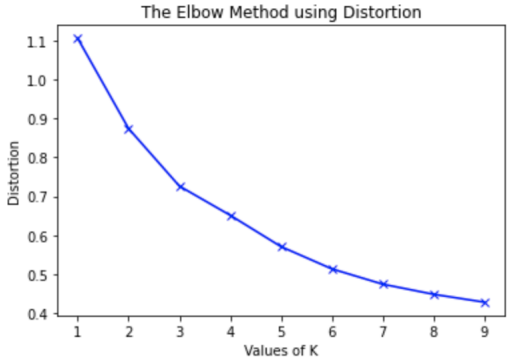

Fig. 6. Optimal count of neighbors

B. Machine learning classifier

The experiment uses 3 stages such as training, validation, and testing to fit the data to the model. Training to validation dataset ratio is 70:30. Around 433 samples are reserved for testing. Euclidean distance metric offered the best confidence score out of different metrics summarized in Fig. 3 and used as the metric in the KNN classifier to classify attack types.

Training data size	615000
Test data size	433
Classifier	KNN Classifier (Euclidean)
Confidence score	0.85

Fig. 7. Data and approach

C. Reinforcement learning

The data will be added to the training set and retrained. This sequence of actions is more of active learning during which similar attacks are identified and used to classify unlabeled data. This phase of reinforcement learning tries to ensure a safe outcome by not letting the ML(AI) agent make undesirable decisions/ classifications.

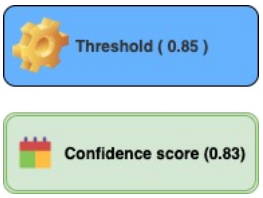

Fig. 8. Threshold and the actual score

D. Result and analysis

The experiment can successfully simulate active/ reinforcement learning with human interference whenever the confidence goes below the threshold level and hence trying to control the outcome of the ML classifier safely by requiring the human to label the suspected low confidence data. This paper is recommending SafeML through Active/ Reinforcement learning to deal with one of the AI safety issues 'Robustness to distributional shift' by trying to minimize the uncertainty of unknown data.

VI. Abbreviations and Acronyms

KNN(K-Nearest Neighbors), SafeML(Safety Machine Learning), AI(Artificial Intelligence), PCA(Principal component analysis)

VII. Conclusion

In a nutshell, with increased autonomy, comes increased risk of the outcome without human intervention. AI-powered applications like robotic surgery, stock trading, political campaigns, self-driving Tesla cars influence the life of mankind to a great deal in the current world. Therefore, addressing the Concrete Problems in AI Safety would help drive AI research and applications in a more proactive direction.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety. arxiv 2016," *arXiv preprint arXiv:1606.06565*.

[2] K. Aslansefat, I. Sorokos, D. Whiting, R. T. Kolagari, and Y. Papadopoulos, "Safeml: Safety monitoring of machine learning classifiers through statistical difference measure," *arXiv preprint arXiv:2005.13166*, 2020.

[3] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge, 1998.

[4] C. Morato, K. N. Kaipa, B. Zhao, and S. K. Gupta, "Toward safe human robot collaboration by using multiple kinects based real-time human tracking," *Journal of Computing and Information Science in Engineering*, vol. 14, no. 1, 2014.

[5] M. A. Cheema, W. Zhang, X. Lin, Y. Zhang, and X. Li, "Continuous reverse k nearest neighbors queries in euclidean space and in spatial networks," *The VLDB Journal*, vol. 21, no. 1, pp. 69–95, 2012.

[6] T. Wei, L.-Z. Guo, Y.-F. Li, and W. Gao, "Learning safe multi-label prediction for weakly labeled data," *Machine Learning*, vol. 107, no. 4, pp. 703–725, 2018.

[7] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization.," in *ICISSP*, pp. 108–116, 2018.