An investigation of Early Cyber Threat Detection using Ensembles of Machine Learning Methods

V. Bulavas, G.Dzemyda, V. Marcinkevičius
Vilnius University Faculty of Mathematics and informatics
Institute of Data Science and Digital Technologies

Vilnius University

Abstract

According to PwC's Global economic crime survey, cybercrime has overall evolved into the second place after an asset misappropriation. According to Lithuanian National Cyber Security Centre Annual report for 2016, scanning of surveilled network devices since 2015 has increased fivefold. Lithuanian academic network LITNET is no different, observing persistent multiple step intrusion activities. As nowadays it is impossible to detect and mitigate all threats manually, automatic tools are used on a 24/7 basis. The techniques utilized by current network intrusion detection appliances in use fall into three main categories: anomaly detection, misuse detection and hybrid. Misuse detection systems use signatures that describe already known attacks and require regular ruleset update. Machine learning based anomaly detection requires supervision and specialist review due to currently still high false positive rate of detecting previously unseen system behaviors. With an increasing frequency of cyber-attack, reviews take more and more time of cyber security specialists, which is a challenge. This indicates highly demanded area for research aiming to increase threat detection accuracy and training speed. Until very recently there was little published research about successful early threat detection models. Other authors proposed an ensemble of Machine Learning models as a probable way for solving abovementioned early detection problems. Therefore, in this work, authors perform investigation of selected method ensembles and present results of comparison.

Keywords

Network security; intrusion detection; early warning; anomaly detection; machine learning; ensemble learning; neuron capsules

Objectives

Response to cyber crime, with an increasing speed becoming everyday hassle to the society and enterprises, requires multi-layered Intrusion Detection Systems (IDS) [Figure 1].

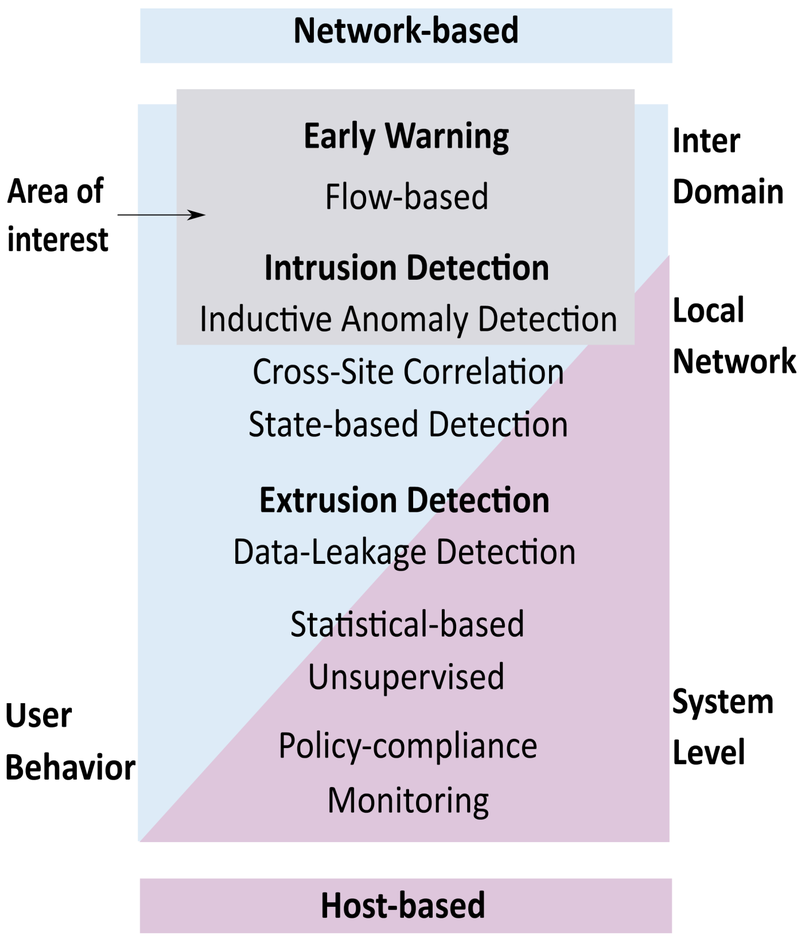

Figure 1. Layers of a Next-Generation IDS [1].

In this research we pay attention to Early Warning layer of network based intrusion detection. The basic style attack types, occurring in cyber space require different types of Machine Learning (ML) algorithm ensembles to reduce the false alert rate (FAR), which varies from 0% to 5% depending on the ML method used and the type of attack. Meanwhile scanning of surveilled network devices is increasing, new threats are addressing diverse multi-layer attack vectors and require intensive use of behavior-based detection techniques. Current objective of research is to investigate the new neuron capsule method for possible application in Intrusion Detection Systems and possibly increase the speed of threat identification.

Issues of Flow-based Detection

The router or switch has the ability to collect IP network traffic as it enters and exits the interface. Flow monitoring has become a prevalent method for monitoring traffic in high-speed networks. A network flow is predominantly defined as a unidirectional sequence of packets that share the exact same packet attributes: ingress interface, source IP address, destination IP address, IP protocol, source port, destination port, and IP type of service. The NetFlow protocol itself has been superseded by Internet Protocol Flow Information eXport (IPFIX), therefore ML training data sets created earlier are becoming obsolete and are far from being universal. Even though NetFlows may be still the most frequent due to Cisco's popularity in the networking industry, other network equipment vendors provide similar network flow monitoring technology, which implies, that flows training has to be tailored for specific equipment on site.

Ensemble Methods for IDS

Ensemble learning methods train combinations of base models traditionally used in supervised learning [Figure 2].

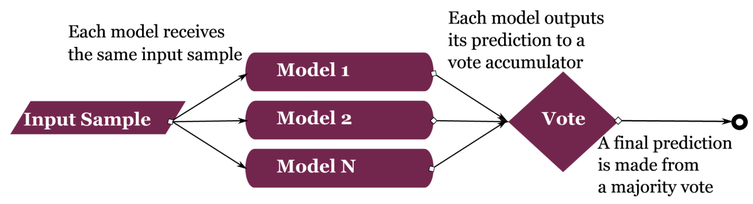

Figure 2. Ensemble of ML Methods.

Machine learning approaches used in intrusion detection include Decision Trees, Inductive Learning, Naive Bayes, Random Forest, Artificial Neural Networks, Fuzzy Systems, Evolutionary Computation, Artificial Immune Systems, Hidden Markov, Sequential Pattern Mining, Swarm Intelligence [2], [3] and other.

Ensembles of ML methods were demonstrated to be an efficient way of improving predictive accuracy and/or decomposing a complex, difficult learning problem into some easier tasks [4]. According to the "no free lunch" theorem, there is not a single classifier that is appropriate for all the tasks, since each algorithm has its own domain of competence. Therefore we need a pool of classifiers to solve a given problem.

However there are several known ensemble learning issues such as the number and types of base models to use, the combining method to use, and how to maintain diversity among the base models. Current IDS require immense amount of data to learn, and data from one source (or location) is not enough. Experts are concerned with a need of constant retraining for IDS and refeeding same data into different models of an Ensemble.

Neuron Capsules

Facing the fact, that network data flows are coming as a stream of virtually never repeating data, the relatively new concept of neuron capsules is expected to help overcoming the limitation of a need of constant retraining. S. Sabour, N. Frosst and G.E. Hinton [5] demonstrated that a discriminatively trained, multi-layer capsule system achieves superior performance, reducing the number of test errors by 45% compared to the previously used ML methods applied in image recognition area.

A capsule is a group of neurons whose outputs represent different properties of the same entity. Active capsules at one level make predictions for the instantiation parameters of higher level capsules. When multiple predictions agree, a higher level capsule becomes active. To achieve these results an iterative routing by agreement mechanism is applied: a lower level capsule prefers to send its output to higher level capsules whose activity vectors have a big scalar product with the prediction coming from the lower level capsule [5].

The Capsule Network is expected to be capable of extracting more understanding from a given amount of data than single Ensemble.

Future Work

This research allows us to substantiate, that if using neuron capsules helped solving classification problem for image data, it is highly probable, that it will be effective with classification of network traffic data.

Therefore the objective of our future research is to investigate the Neuron Capsule method application for early warning and intrusion detection tasks.

For that we need to build a test environment, train the system and calculate indicators for comparison of classical Ensemble methods and the new Capsule Networks.

Literature

1. Koch, R. (2011). *Towards Next-Generation Intrusion Detection*. In T. W. (Eds. C. Czosseck, E. Tyugu (Ed.), 2011 3rd International Conference on Cyber Conflict (pp. 151–168). Tallinn, Estonia: CCD COE Publications.
2. F. Gharibian and A. Ghorbani (2007). *Comparative study of supervised machine learning techniques for intrusion detection*, in Proc. 5th Annu. Conf. Commun. Netw. Serv. Res., pp. 350–358.
3. Buczak, A., & Guven, E. (2015). *A survey of data mining and machine learning methods for cyber security intrusion detection*. IEEE Communications Surveys & Tutorials, pp. (1153-1176)
4. Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). *Ensemble learning for data stream analysis: A survey*. Information Fusion, 37, 132–156.
5. Sabour, S., Frosst, N., & Hinton, G. E. (2017). *Dynamic Routing Between Capsules*. Retrieved from http://arxiv.org/abs/1710.09829