Temasek Defence Systems Institute

Validation of Machine Learning Algorithm on the Intrusion Detection System (IDS) of Navy Smart Grid

ME5 Ng Wee San, Republic of Singapore Air Force
Prof. Preetha Thulasiraman, ECE Dept., Naval Postgraduate School

Introduction

In 2013, the U.S. Naval Facilities Engineering Systems Command (NAVFAC) developed the operational requirements to implement its own smart grid [1]. In 2019, NAVFAC has achieved Full Operational Capability (FOC) of the smart grid at several locations in the mid-Atlantic states. For the Navy to ensure resiliency of the smart grid network, effective countermeasures against security threats must be addressed using techniques that adapt to the massive data that is collected by smart sensors and meters.

Objectives

The goal of our research is to develop cyber threat detection techniques to secure energy infrastructure that is critical to the Navy. Our research is focused on the use of cyber analytics to continually define and mitigate evolving threat vectors for the Navy Smart Grid. The contributions of this thesis are:

- Design of an IDS architecture for the Navy smart grid.
- Evaluation and comparison of KNN, Bayesian and Random Forest algorithm on Lemay and Fernandez's datasets.
- Validation via MATLAB that the Random Forest algorithm is a more efficient approach based on performances shown on true positive rates, Matthew Correlation Coefficient (MCC) values and overall accuracy.

Machine Learning Algorithms

K-Nearest Neighbour. KNN uses Euclidean metric to quantify the differences between each sampled data. Euclidean distance is described as follows:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} w_r (a_r(x_i) - a_r(x_j))^2}$$

where we define vector $x = (a_1, a_2, a_3, \dots, a_n)$, where *n* denotes the vector input's dimensionality, or the number of sample characteristics. a_r is the example's *r*th characteristics while w_r is the weight of the *r*th characteristics. *r* ranges from *1* to *n* [2]. This means that the smaller the Euclidean distance, $d(x_i, x_j)$ between any two examples, the similarity between them rises.

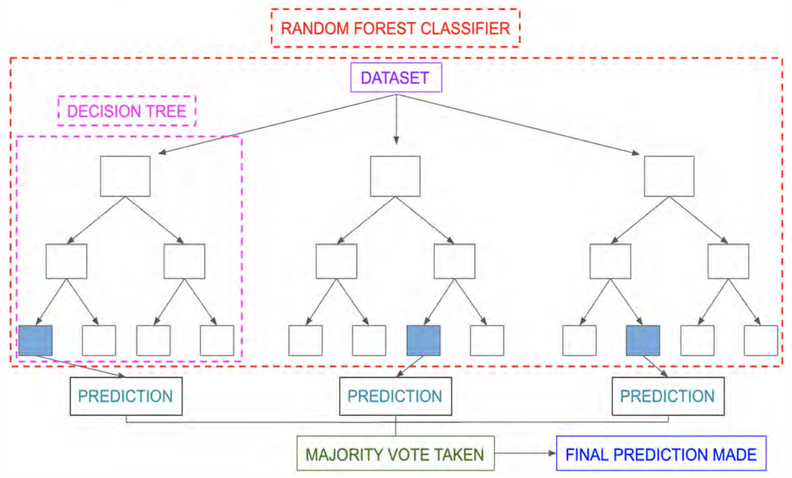
Bayesian. This algorithm comprises directed acyclic graphs (DAGs) of a parent node with several child nodes where the child nodes are independent of one another. The categorization is guaranteed by deliberating the parent node as a concealed parameter that specifies which class each item in the dataset should be allocated to, and the child nodes as various properties that define this object. The Bayesian rule is expressed as such:

$$P(s_i \mid E) = \frac{P(E \mid s_i) \cdot P(s_i)}{P(E)}$$

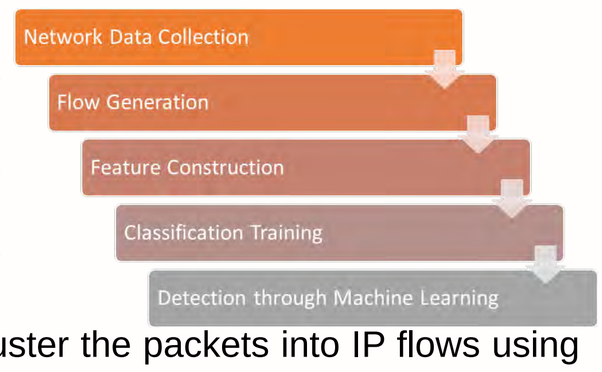
where in the session class, E denotes the sum of evidence on attribute nodes and s_i represent a potential value. The sum of evidence E are dispersed into individual parts., say e_1, e_2, ..., e_n with relation to E_1, E_2, ..., E_n, respectively [3].

Random Forest. Random Forest is an ensemble of unpruned classification or regression trees usually trained with the 'bagging' method [4]. The algorithm produces various classification trees constructed from decision tree algorithms.

Each decision tree is comprised of a data sample taken from a training set, called the bootstrap sample. Randomness is induced through feature bagging to reduce the correlation among all the classification trees. A majority vote based on the most frequent categorical variable of the classification trees will be the final prediction for the algorithm. Figure below depicts a simplified Random Forest classifier.



Proposed Intrusion Detection System (IDS)



An overview of the proposed IDS is shown in the figure on the right. The first step of our IDS architecture is Network Data Collection. This process is responsible for collecting and capturing the Modbus/TCP data. The second step, Flow Generation is where the packet data extracted is used to cluster the packets into IP flows using an open-sourced software called CICFlowMeter. The next step is to construct relevant features where features are extracted, and data instances are generated. The data instances will next be fed to the input of the machine learning algorithm to train and classify the data into malicious or normal traffic. Lastly, the trained machine learning algorithm is used to detect and inform the administrator for the next course of action.

Results

Using MATLAB's machine learning toolbox, we showed the overall accuracy for the simulation runs for the three machine learning algorithms with Random Forest performs the best but slightly better than KNN.

Experiments	Random Forest	KNN	Bayesian
Normal - Malicious	99.69%	99.53%	99.37%
Normal – Exploit – Fingerprint – Malware – Unauthorize	99.53%	99.47%	95.95%

However, the overall accuracy rate may not be a reliable parameter for unbalanced dataset as it does not accurately represent how malicious packets are classified. Thus, we also considered additional measures such as MCC to provide additional insight in the classifier performances.

	Random Forest	KNN	Bayesian
MCC	0.77	0.63	0.47

Results based on MCC values indicate that the random forest implementation performs better than the other two classifiers.

Future Work

We conducted our simulation based on publicly available dataset and using supervised learning methods in our study. To that end, following are recommended for future work:

1. Generate datasets from Navy Smart Grid.
2. Using unsupervised machine learning algorithms

References:
[1] "Navy and Marine Corps Smart Grid CDD Industry Version," NAVFAC, Washington, DC, USA, 2014.
[2] S. Sun and R. Huang, "An Adaptive k-Nearest Neighbor Algorithm," in 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Shanghai, China, 2010.
[3] F. V. Jensen, "Introduction to Bayesian Networks," UCL Press, 1996.
[4] N. B. Amor, S. Benferhat and Z. Elouedi, "Naive Bayes vs Decision Trees in Intrusion Detection Systems," in ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004.

NUS National University of Singapore