Combining Machine Learning and Behavior Analysis Techniques for Network Security

1st Maria Laura Brzezinski Meyer
Double degree student by CAPES/Brafitec program
Federal University of Santa Catarina
Florianopolis, Brasil
ENSEEIHT INP-Toulouse
Toulouse, France
laurabrzmeyer@gmail.com

2nd Yann Labit
Université de Toulouse
Laboratoire d'analyse et d'architecture des systèmes - CNRS
Toulouse, France
ylabit@laas.fr

***Abstract*—Network traffic attacks are increasingly common and varied, this is a big problem especially when the target network is centralized. The creation of IDS (Intrusion Detection Systems) capable of detecting various types of attacks is necessary. Machine learning algorithms are widely used in the classification of data, bringing a good result in the area of computer networks. In addition, the analysis of entropy and distance between data sets are also very effective in detecting anomalies. However, each technique has its limitations, so this work aims to study their combination in order to improve their performance and create a new intrusion detection system capable of well detect some of the most common attacks. Reliability indices will be used as metrics to the combination decision and they will be updated in each new dataset according to the decision made earlier.**

***Index Terms*—Combination, IDS, Security, Neural Network, Decision Tree, Random Forest, Hausdorff Distance, Kullback-Leibler Divergence, Entropy, Network Attacks, DoS, Confidence Index.**

I. Introduction

The great technological advance of nowadays brings many facilities and economy of time, however, this expansion causes security problems to interconnected machines and networks. To overcome this problem, intrusion detection systems (IDS) are implemented, which are devices capable of tracking packages headers and data to detect invasions and accesses that may degrade network performance. This project aims to train an intrusion detection algorithm using machine learning techniques and the system's behavior analysis. The considered attacks are Denial of Service (DoS), Probing (Proba), Remote to Local (R2L), User to Root (U2R).

Machine learning techniques are widely used for outlier detection or data classification. Borji [1] uses three combination methods (belief function, majority voting and Bayesian average) to arrange four machine learning classifiers proving that it achieves a better result than each algorithm individually. According to Zouari et al. [2], there are two types of combination methods: when multiple results from the same classifier are used (for example boosting, bagging) and when different algorithms are combined (using supervised or unsupervised algorithms). The relative entropy theory is also used to detect network anomalies, like in the work of Zhang et al. [3]. Even that the detection rates of each attack obtained are low, the Kullback-Leibler divergence is a reasonable parameter to analyze the traffic behavior. But it has a strong dependency on the selected features for the composition of the probability distribution. The simple entropy is used for traffic analysis with and without attacks in Hamamoto et al. [4], showing that this parameter changes visibly in the presence of attacks for some features and can be also a parameter for behavior analysis.

In this project, both machine learning predictions (detection by signature) and behavior analysis are used (detection by pattern). So two combinations are done: the first is fusion of four algorithms predictions; then, in the second one, the prediction must be in accordance with traffic behavior analysis. Confidence indices are used in decision making and updated with each new analysis, so a better accuracy can be achieved. The method will be more detailed in section 2, as well as the machine learning techniques and the entropy, Kullback-Leibler divergence and Hausdorff distance calculations. In section 3, the test environment is described and the results are shown. Finally, in section 4, the conclusions are drawn and future works are suggested.

II. Method

The proposed intrusion detection system aims to train an attack detection algorithm through the system pattern. The behavior is analyzed using the Shannon Entropy, Hausdorff distance, and Kullback-Leibler divergence. Four machine learning algorithms predictions are combined and then compared to behavior parameters to confirm that there is an anomaly in the sample. Then the decision is stocked to be used in the behavior algorithm's train. An overview is shown in figure 1, where five processes are shown: the data treatment, the separation of windowed packages to be analyzed with each interaction, the machine learning algorithms, the decisions of each parameter (Entropy, Hausdorff, Kullback-Leibler and Machine Learning), the final decision and the parameters actualization.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)

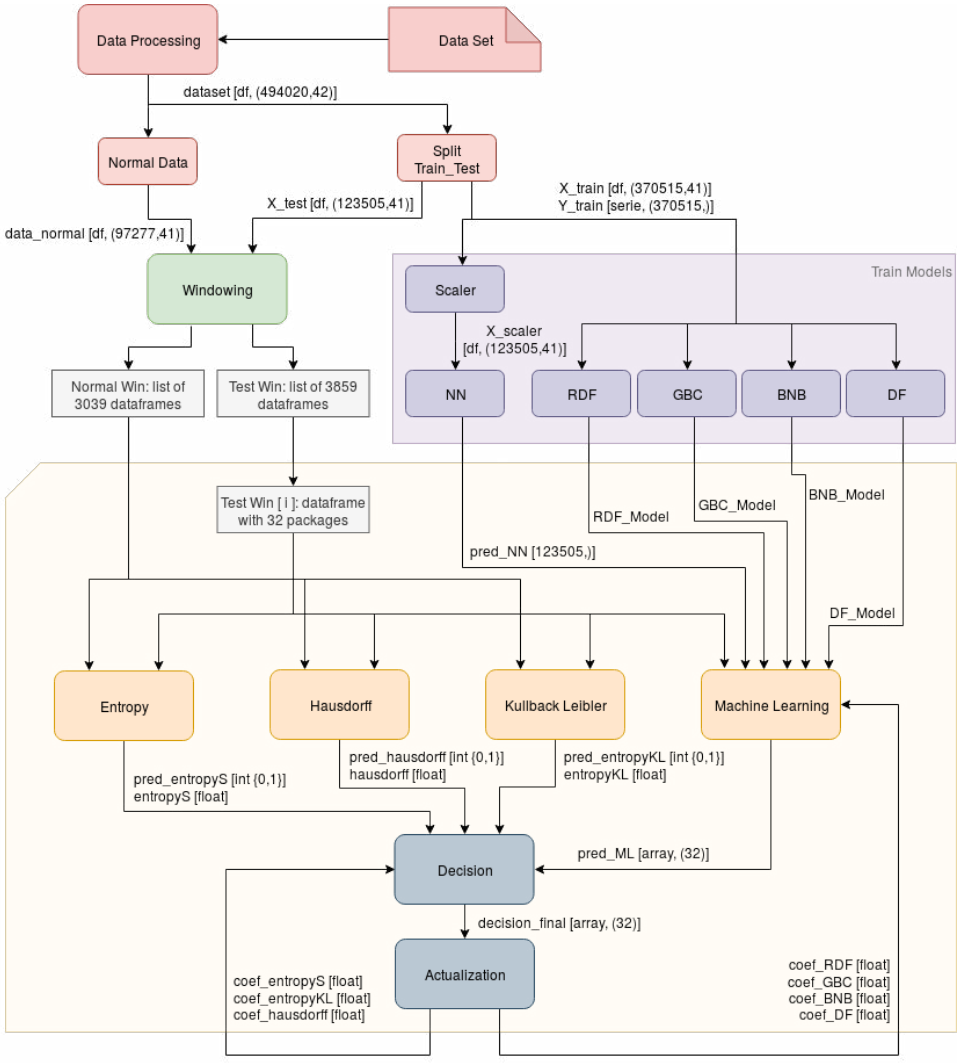

Fig. 1. Method overview

A. Machine Learning

The machine learning algorithms were tested using 70% of the KDD CUP 99 dataset [8] to training and 30% for testing, the tests results are shown in figure 2. Taking into account the train time of the tested algorithms and their performance, four machine learning methods were selected. The first one is Bernoulli Naive Bayes (BNB) because it takes little training time and has a good detection of normal and DoS traffic. Decision Tree (DT) and Random Forest (RDF) are also chosen because of their good detection of Proba, R2L and U2R traffic. The last is Gradient Boosting (GBC) that requires an acceptable training time and has good accuracy comparing to the others. The system took one algorithm as the most trustworthy that is used as a reference to construct the other algorithms confusion matrices, which is necessary to calculate the confidence indices of each classifier allowing the combination of their predictions. As KNN requires a lot of training time, that is why the second best performance algorithm - neural network (MLP NN) - was chosen to estimate the labels.

The combination of this four algorithms is made using a confidence index for each algorithm prediction. So if an algorithm prediction is considered wrong compared to the result of the others, a punishment is applied by decreasing the confidence index of this algorithm. Similarly, an increase is made when the algorithm is considered reliable.

B. Entropy Simple

The entropy is a measure of a system's irreversibility, it is usually associated with the system's disorder. The Shannon entropy, or simple entropy, measure the unlikelihood of a given event in a probability distribution and can be calculated as below, where P is the probability mass function for a possible value x_i of $X = x_1, x_2, ..., x_n$:

$$H(x) = -\sum_{i=1}^{n} P(x_i) log P(x_i) \quad (1)$$

	Accuracy	F1-Score	Recall	Time Training (s)
KNN	99.882	83.066	80.559	514.623111
NN	99.675	75.964	73.693	230.776599
SVM	95.507	39.644	37.722	7274.630503
LR	99.388	65.321	62.474	712.796136
GNB	79.972	39.666	60.709	0.289856
BNB	92.774	59.877	67.476	0.198194
DT	99.965	91.719	90.153	1.090529
RDF	99.979	90.978	87.443	4.344582
ADB	85.621	44.558	56.113	29.642418
GBC	99.940	82.784	81.121	215.867362
SGD	92.660	40.206	40.680	21.005275

Fig. 2. Machine learning algorithms performance

J. S. Paz and D. T. Roman address the use of entropy to detect anomalies in [5]. The entropy of normal traffic windows are calculated to fix a threshold, so the current traffic entropy can be compared to this criterion. For the simple entropy analysis, the probability distribution of the data bytes source-destination number is considered. When two simple entropy of source bytes are compared, the traffic is abnormal if its entropy is below a threshold. This is because there is no standard data bytes number in normal traffic, resulting in high entropy. Then when an attack is present, its packages normalize the number of data bytes, which causes the entropy to decrease.

C. Kullback-Leibler Divergence

The Kullback-Leibler divergence, as known as relative entropy, is also utilized to analyze traffic behavior, like in [6]. This parameter measures how much different a probability distribution $P(i)$ is from another $Q(i)$ and it is described as, where i represents a discrete random variable:

$$D_{KL}(P||Q) = -\sum_{i} P(i) log \frac{Q(i)}{P(i)} \quad (2)$$

A threshold is calculated using normal traffic windows and then, the divergence between the current traffic and a normal one is compared to this criterion. If the divergence is higher than the threshold, an anomaly is present, otherwise, the traffic is normal.

D. Hausdorff distance

The Hausdorff distance measures how distant one data set X is from another Y and it can be also as a metric to distinguish abnormal from normal traffic [7]. It is mathematically represented as:

$$d_H(X,Y) = max\{\underset{x\in X}{sup}\ \underset{y\in Y}{inf}\ d(x,y), \underset{y\in Y}{sup}\ \underset{x\in X}{inf}\ d(x,y)\} \quad (3)$$

Where sup represents the supremum (the least upper bound), inf the infimum (the greatest lower bound), x is a data from X and y is a data from Y. It is used to know the distance between the current traffic and normal traffic. To do so, a threshold is fixed based on normal traffics windows and then, if it is higher then the distance between the current traffic and a normal one, the traffic is labeled as abnormal, otherwise, it is a normal traffic.

E. Final Decision

Fist of all, all methods are used to analyse the data. The behavior methods (simple entropy, Kullback-Leibler divergence and Hausdorff distance) can distinguish abnormal traffic from normal traffic, but they can't tell witch type of attack is present. That's why machine learning algorithms are needed, they are combined to better classify the attacks types. This combined prediction is compared to the analysis of the behavior to confirm that an anomaly is present or not, then a final decision is made. If there are anomalies, the machine learning classification and the traffic parameters are stocked in a data frame as shown in the figure 3. This data is used to train the behavior algorithm, that searches for a similar set of parameters already observed to make its own prediction, so they learn to distinguish the different types of attacks. Therefore, with each interaction, behavior analysis becomes less dependent on machine learning analysis.

	Entropy	KLD	Hausdorff	Normal	DoS	Proba	R2L	U2L
45	1.946466	9.930987	53835.356890	1.0	1.0	0.0	1.0	0.0
2	1.984775	11.006686	90892.672215	1.0	1.0	0.0	0.0	0.0
217	1.380088	12.105585	53673.627670	1.0	1.0	0.0	0.0	1.0
3113	1.414234	10.696530	55813.729138	1.0	1.0	0.0	1.0	0.0
24	1.481228	13.726850	57155.327203	1.0	1.0	1.0	0.0	0.0

Fig. 3. Example of data behavior and prediction

III. First Results

All the tests were made in a Ubuntu 16.04 virtual machine with 16384 Mb of memory and 4 processors. The Anaconda environment was used to create Jupyter notebooks in Python 3. The KDD CUP 99 data set [8] is the most employed data set for network security analyses, so 10% of it is loaded from the UCI Machine Learning Repository [1] to be used in the tests.

A test using 3859 windows with 32 packages each was made and the score of the final decision predictions was 99.97%. This number refers to the percentage of packets correctly classified by the intrusion detection system using a neural network predictions as a reference. This testing data is formed by 24388 packages of normal traffic, 97788 of DoS attack, 1035 of probing attacks, 262 of R2L attacks and 15 of U2R attacks. The behavior algorithm's performance is shown in figure 4, where the *Count* column represents how much windows have the specified type of traffic. In the end there are two predictions, one from the intrusion detection system and other form the trained algorithm. The idea is that, after a

[1]https://datahub.io/machine-learning/kddcup99

certain number of processed data, machine learning algorithms will no longer be needed.

	Accuracy	Count
Normal	98.2638	3854
DoS	98.3934	3859
Proba	76.3410	909
R2L	94.0658	242
U2R	99.7409	10
Total	99.9700	3859

Fig. 4. Algorithm prediction's accuracy

Using four machine learning algorithms in [1], A. Borji achieved a detection rate of 99.18% for majority voting combining, 99.3% using a Bayesian combining and 99.68% using belief combining. Using a fuzzy combination of artificial neural network algorithms, G. Wang et al. obtained an average of 96.71% in [9]. With the goal of selecting an optimal attribute set for a better classification of attacks, in [10] tree entropy types are compared. Using the Shannon one and K-means algorithm, they obtained an accuracy of 70.10% for DoS attacks, 66.97% for probing, 47.00% for R2L and 47.87% for U2R. They also combined the entropy analysis with the Farthest First Traversal Algorithm by Hochbaum and Shmoys [11], which resulted in better detection of DoS (90.86%), R2L (95.53%) and U2R (95.05%) attacks, however for probing attacks the accuracy was lower. All the results of these articles were also based on the KDD CUP 99 data set.

IV. Conclusion

The Intrusion Detection System created using machine learning and behavior metrics (entropy, Kullback-Leibler divergence and Hausdorff difference) combination, as well as reliability indices, has a good result. The algorithm can achieve a reliable decision after a certain number of interactions without the support of machine learning algorithms. Although the flawed in detecting probing attacks, the algorithm's results are promising and adjustments can be made to improve it in this regard. The final decision has a great accuracy rate in comparing to the performance of machine learning algorithms separately. Future work will focus on testing unsupervised machine learning algorithms, as well as using other data sets - such as NSL KDD [12] and DARPA [2] - to verify the intrusion detection system's adaptability.

The Dempster–Shafer theory, developed by Arthur P. Dempster in *Upper and lower probabilities induced by a multivalued mapping* (1967) and Glenn Shafer in *A Mathematical Theory of Evidence* (1976), allows combining evidence from different classifiers to achieve a degree of belief for each one. This index can be used to improve the machine learning algorithms combination. This theory is used in data classification and the terms degree of plausibility and degree of belief are described in the paper [13]. This is one of the methods used to combine the classifiers in [1]. Surathong et al. [14] make a fuzzification of the belief mass to improve the combined decision and Xu et al. [15] applied the theory to combine multiple classifier's decisions in handwriting recognition. They all have interesting results using the belief function as a combination parameter.

[2]https://www.ll.mit.edu/r-d/datasets

Acknowledgment

The authors would like to thank research laboratory LAAS (*Laboratoire d'analyse et d'architecture des systèmes*) for the resources necessary and also the brazilian university UFSC (*Universidade Federal de Santa Catarina*), the french higher school ENSEEIHT (*École Nationale Supérieure d'Électrotechnique, d'Électronique, d'Informatique, d'Hydraulique et des Télécommunications*) and the Toulouse University. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

[1] A. Borji, "Combining Heterogeneous Classifiers for Network Intrusion Detection", Advances in Computer Science – ASIAN 2007. Computer and Network Security, 254–260, 2007.

[2] H. Zouari, L. Heutte, Y. Lecourtier and A. ALimi, "Un panorama des méthodes de combinaison de classifieurs en reconnaissance de formes", (French) [*An overview of classifier combination methods in pattern recognition*]. Proc. RFIA'2002, Angers, France, vol.2, 499–508, 2002.

[3] Y. Zhang, Z. Han and J. Ren, "A Network Anomaly Detection Method Based on Relative Entropy Theory", Second International Symposium on Electronic Commerce and Security, 231–235, 2009.

[4] A. H. Hamamoto, A. Carvalho, L. D. H. Sampaio, L. Abrao, T. L. P. Mario, "Network Anomaly Detection System using Genetic Algorithm and Fuzzy Logic", Expert Systems with Applications, vol.92, 390-402, February 2018.

[5] J. S. Paz and D. T. Roman, "On Entropy in Network Traffic Anomaly Detection", 2nd International Electronic Conference on Entropy and Its Applications, 10.3390/ecea-2-B008, 2015.

[6] C. Maklin, "KL Divergence Python Example", Towards Data Science, `https://towardsdatascience.com`, accessed in July 2019.

[7] Y. Labit and J. Mazel, "HIDDeN: Hausdorff distance based Intrusion Detection approach DEdicated to Networks", The Third International Conference on Internet Monitoring and Protection, 11–16, 2008.

[8] D. Dua and C. Graff, "UCI Machine Learning Repository", Irvine, CA: University of California, School of Information, accessed in July 2019.

[9] G. Wang, J. Hao, J. Ma and L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering", Expert Systems with Applications, vol.37, 6225–6232, September 2010.

[10] C. F. L. Lima, F. M. Assis and C. P. de Souza, "A comparative study of use of Shannon, Rényi and Tsallis Entropy for Attribute Selecting in Network Intrusion Detection", Springer Berlin Heidelberg, Intelligent Data Engineering and Automated Learning, 492–501, 2012.

[11] D. S. Hochbaum and D.B. Shmoys, "A best possible heuristic for the k-center problem", Mathematics of Operations Research, vol. 10, no. 2, pp. 180–184, 1985.

[12] M. Tavallaee, E. Bagheri, W. Lu and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", Second IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.

[13] Q. Chen, A. Whitbrook, Uwe Aickelin and Chris Roadknight, "Data Classification Using the Dempster-Shafer Method", Journal of Experimental Theoretical Artificial Intelligence, 2014.

[14] S. Surathong, S. Auephanwiriyakul and N. T. Umpon, "Decision Fusion Using Fuzzy Dempster-Shafer Theory", Recent Advances in Information and Communication Technology, 115–125, 2018.

[15] L. Xu, A. Krzyzak and C. Y. Suen, "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition", IEEE Transactions on Systems, Man, and Cybernetics, vol.22, no.3, 418–435, May-June 1992. DOI: 10.1109/21.155943