Prediction of attacks types in telecommunications network with machine learning

Mureille Laure Obaya*
Tutor and Hope project manager
Be Educated and Learn Skills in Artificial Intelligence (BEL'S AI)
Yaounde, Cameroon
belsaiinitiative.org, laureoby@gmail.com

1 Abstract

Telecommunication networks are essential for the development of today's societies. In professional, social and other relationships, it is almost impossible to live without networks. To ensure the safety of users, it is important to protect them from potential attacks that are likely to occur. It is for this reason that the Intrusion Detection System (IDS) has been set up. With to machine learning, we decided to create a model allowing us to predict the types of attacks in a telecommunications network. For this work we used the NSK KDD dataset.

1. 2. Context

Data mining and machine learning have been the subject of much research in the detection of intrusions in a telecommunications network. The NSL KDD dataset is the most widely used dataset in research. Intrusion detection can be performed on a machine (HIDS – Host-based Intrusion Detection System) or on a network (NIDS – Network-based Intrusion Detection System). IDSs are also categorized according to the method used: by anomaly or by signature. The latter is the most commonly used method.[1]

Machine learning algorithms learn autonomously to perform a task or make predictions from data and improve their performance over time[2]. Once trained, the algorithm will be able to find patterns in new data. The algorithm typically learns network traffic on a training dataset in a supervised manner, which may contain attacks. Each data in the dataset is labeled as normal or as an attack. Once trained, this algorithm is used on a test dataset to evaluate its performance on traffic it has never encountered before.

To better develop our theme, we will first present the dataset that we used in our work. Then, we will talk about the different stages of processing our data then the training and testing phase and we will end this paper with a conclusion and perspectives.

3. The studied dataset

For this work, we use the NSL-KDD intrusion detection dataset. It is based on another popular dataset, the KDD Cup 99. The latter was created in 1999 for a machine learning competition [3]. The goal of this competition was to correctly classify network connections into 5 categories:

normal, denial of service (DoS), network probe (probe), remote to local (R2L – Remote to Local), user to root (U2R – User to root). Each connection has 41 characteristics that allow the classifier to correctly predict its class. These characteristics are information or statistics calculated from listening to a simulated local area network of the U.S. Air Force in 1998: duration of connection, type of protocol, percentage of connections to the same service, etc.

NSL-KDD was created in 2009 to solve some problems inherent in KDD Cup 99 [4]. He thus uses the same data as the latter, but modifies it greatly to make his corrections. Thus, redundant or duplicate connections, which made up 75% to 78% of the dataset, were removed. The total number of data is greatly reduced: it goes from 805050 connections for KDD Cup99 to 148517 for NSL-KDD.

KDD Cup 99 has been repeatedly criticized by the scientific community for the issues we have discussed [4]. Although NSL-KDD corrects some of it, it is based on the same data that dates back to 1998. It is therefore far too old to be a reliable representation of current network activity and threats. It is also worth noting the strong disparity in the distribution of NSL-KDD classes between training and test sets.

In the field of artificial intelligence, Machine Learning is the technology that allows machines to learn on their own from data provided with the aim of making the machine or computer capable of providing solutions to complex problems by processing an astronomical amount of information. This thus offers an opportunity to analyze and highlight the correlations that exist between two or more given situations, and to predict their different implications. [3] It can be used in several fields: the prediction of financial values, the detection of intrusion in the field of computer security, the search engine influenced by the profile of the user, the detection of machine thefts, the implementation of an anti-virus and cryptanalysis. The implementation of machine learning takes place according to the following steps:

- Obtaining and cleaning data
- Realization of the model
- Learning phase
- Validation phase
- Execution phase

4. Processing of dataset

2. 4.1 Data discretization and Data transformation

In the dataset, we detected four attributes with non-numeric values. We have digitized them (DATA DISCRETIZATION) by replacing the given character string values into precise integers. These numbers will be between [0;n] (DATA TRANSFORMATION). These attributes are: "tcp", "normal", "ftp_data" and "SF".

3. 4.2 Data cleaning

We highlighted all the missing values of our dataset to then clean it in case they exist in a record. We also brought out redundant data to clean them up. Note that the missing values are most often of the "Null", "NaN" or "None" type.

4. 4.3 Data reduction

5. Using several graphs, we have analyzed the attributes in order to select those which give the most information to facilitate the prediction of the type of attacks. This allowed us to remove the attributes that are the least correlated with the prediction variable: "20". This step marks the end of the first phase of our data processing. We have a new dataset that we worked with in the training phase of the model.

5. Learning and testing phase

We have divided the dataset into training data and test data. To do this, we split the dataset into two global variables, one representing the output variable of dataset Y and the other representing the input variables of dataset X. We then, split the dataset into training dataset(x_train, y_train) and test dataset(x_test, y_test). To train our model, we used a machine learning model adapted to our problem which is a classification algorithm. For this, we used the MLPClassifier model from the sklearn.neural_network library by creating a function to test and evaluate the model and by applying this function to our dataset.

After applying our algorithm, we had to choose the appropriate model to design and develop a classification algorithm to perform the prediction of attacks in a telecommunication network. We have chosen our model among the following models:

- Logistic Regression
- Gaussian Naive Bayes
- SVM (NEURAL NETWORK)

Of these 03 models, the one we judged to be more efficient is the Logistic regression whose learning curve is presented below:

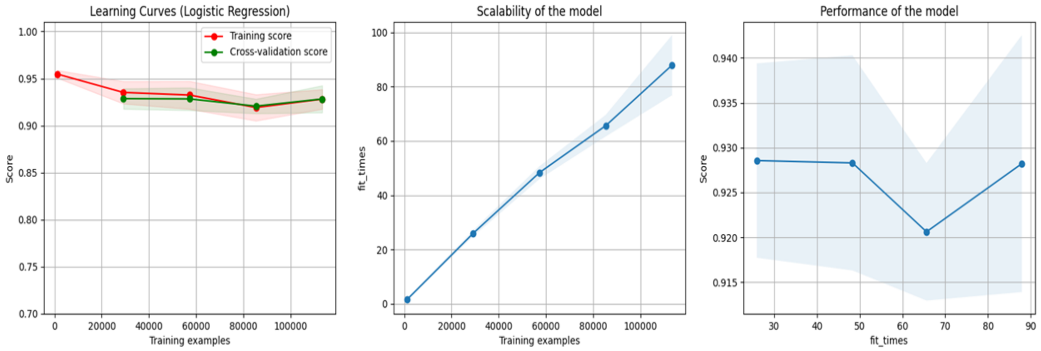

Figure 1: Learning curve Logistic regression

6. Conclusion

By way of conclusion, we can say that IDSs are important assets that make it possible to best ensure the security of users in a telecommunications network. There are different types of attacks and machine learning can effectively help detect them and prevent their effect on the proper functioning of a network.

We would like to apply the algorithm we have chosen on a more recent dataset. In addition, we would like to go further by applying the model chosen for the realization of a viable solution that can be used by companies or ordinary users.

References

[1] Douid Rania, 2019. Système de détection d’intrusion réseau basé sur L’algorithme de Classification KNN, *Projet de fin d’etude pour l’obtention du diplome de master en securite des systemes d’information*, universite saad dahlab de blida 1, Algérie. DOI: https://di.univ.blida.dz/jspui/handle/123456789/4069

[2] https://datascientest.com/machine-learning-tout-savoir, consulté le 29 juillet 2023.

[3] Maxime LABONNE, Alexis OLIVEREAU & Djamal ZEGHLACHE. 2018. Automatisation du processus d’entraînement d’un ensemble d’algorithmes de machine learning optimisés pour la détection d’intrusion. In *Cesar conference, Artificial Intelligence and Cybersecurity*. Rennes, France. DOI:https://www.cesar-conference.org/wp-content/uploads/2018/11/articles/C&ESAR_2018_J2-09_M-LABONNE

[4] M. Tavallaee, E. Bagheri, W. Lu, et A. A. Ghorbani, « A detailed analysis of the KDD CUP 99 data set », in Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on, 2009. DOI:https://ieeexplore.ieee.org/document/5356528