Python Libraries and Entities for Machine learning

E Nagalakshmi Reddy[1], J.S. Ananda Kumar[2]
[1]PG Scholar, Dept. of MCA, Sietk, Puttur,
[2] Assistant Professor, Deptof MCA, Sietk, Puttur, A.P.

ABSTRACT

In order to ensure a company's Internet security, SIEM (Security Information and Event Management) system is in place to simplify the various preventive technologies and flag alerts for security events. Inspectors (SOC) investigate warnings to determine if this is true or not. However, the number of warnings in general is wrong with the majority and is more than the ability of SCO to handle all awareness. Because of this, malicious possibility. Attacks and compromised hosts may be wrong. Machine learning is a possible approach to improving the wrong positive rate and improving the productivity of SOC analysts. In this article, we create a user-centric engineer learning framework for the Internet Safety Functional Center in the real organizational context. We discuss regular data sources in SOC, their work flow, and how to process this data and create an effective machine learning system. This article is aimed at two groups of readers. The first group is intelligent researchers who have no knowledge of data scientists or computer safety fields but who engineer should develop machine learning systems for machine safety. The second groups of visitors are Internet security practitioners that have deep knowledge and expertise in Cyber Security, but do Machine learning experiences do not exist and I'd like to create one by themselves. At the end of the paper, we use the account as an example to demonstrate full steps from data collection, label creation, feature engineering, machine learning algorithm and sample performance evaluations using the computer built in the SOC production of Seyondike.

Keywords: Machine learning, Performance, Stability, Easy of developing, Secure

INTRODUCTION

Cyber security incidents will cause significant financial and reputation impacts on enterprise. In order to detect malicious activities, the SIEM (Security Information and Event Management) system is built in companies or government. The system correlates event logs from endpoint, firewalls, Intrusion Detection/Prevention System), DLP (Data Loss Protection), DNS (Domain Name System), Dynamic Host Configuration Protocol, Windows/Unix security events, VPN logs etc. The security events can be grouped into different categories. The logs have terabytes of data each day. From the security event logs, SOC (Security Operation Center) team develops so-called use cases with a pre-determined severity based on the analyst's experiences. They are typically rule based correlating one or more indicators from different logs.

These rules can be network/host based or time/frequency based. If any pre-defined use case is triggered, SIEM system will generate an alert in real time. SOC analysts will then investigate the alerts to decide whether the user related to the alert is risky (a true positive) or not (false positive). If they find the alerts to be suspicious from the analysis, SOC analysts will create OTRS (Open Source Ticket Request System) tickets. After initial investigation, certain OTRS tickets will be escalated to tier 2 investigation system (e.g., Co3 System) as severe security incidents for further investigation and remediation by Incident Response Team. However, SIEM typically generates a lot of the alerts, but with a very high false positive rate. The number of alerts per day can be hundreds of thousands, much more than the capacity for the SOC to investigate all of them. Because of this, SOC may choose to investigate only the alerts with high severity or suppress the same type of alerts. This could potentially miss some severe attacks. Consequently, a more intelligent and automatic system is required to identify risky users.

Significance of Python

The significance of python is explained as below:

- ➢ python is high –level language
- ➢ python is a both procedure and object – oriented language
- ➢ python is a portable language
- ➢ python is an interactive programming language
- ➢ python is an extendable programming language
- ➢ python is having extensive array of libraries
- ➢ python is a dynamically typed programming language

Significance of Machine Learning

The significance of machine learning is described as follows:

- High data processing
- It is need for feature engineering
- Best results with unstructured data
- No need for labelling data
- It is delivering high quality products with efficient
- Easy of development
- Head – head to model competitions

Python Libraries and Packages for Machine Learning:

Python Librariesfor Data Analysis:

Pandas:

Pandas are becoming to be the most popular Python library that is used for data analysis with support for fast adaptable, and expressive data structures designed to work on both “relational” or “labeled” data. Pandas today is a necessary library for solving practical, real-world data analysis in Python.

Numpy:

NumPy is a well-known general-purpose array-processing package. A sizeable collection of high complexity mathematical functions makes NumPy powerful to process large multi-dimensional arrays and matrices. NumPy is very helpfull for handling linear algebra, Fourier transforms, and random numbers.

Spicy:

The spicy library suggestion modules for linear algebra, image optimization, integration interpolation, special functions, Fast Fourier transform, signal and image processing, Ordinary Differential Equation (ODE) solving, and other computational tasks in science and analytics. The primary data structure used by Spicy is a multi-dimensional array provided by the NumPy module. SciPy count on NumPy for the array manipulation subroutines. The SciPy library was made to work with NumPy arrays together with providing user-friendly and efficient numerical functions.

Stats models:

Stats Models Python package is the best forway creating statistical models, data handling and model evaluation. in addition, with using NumPy arrays and scientific models from SciPy library, it also integrates with Pandas for effective data handling. This library is closely known for statistical computations, statistical testing, and data exploration.

Python Libraries for Data Visualization:

Matplotlib:

Matplotlib is a data visualization library that is helpfull for 2D plotting to produce publication-quality image plots and figures in a variety of formats. The library helps to achieve histograms, plots, error charts, scatter plots, bar charts with just a few lines of code.

Seaborn:

The Matplotlib library design the base of the Seaborn library. In similarity to Matplotlib, Seaborn can be used to create more appealing and descriptive statistical graphs. Along with large-scale supports for data visualization, Seaborn also comes with an inbuilt data set oriented API for studying the relationships between multiple variables.

Python Libraries for Machine Learning:

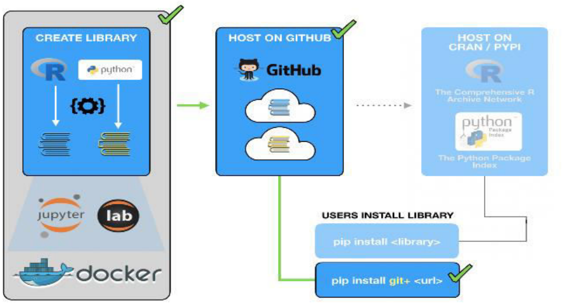
Scikit – learn:

Scikit-learn has a wide range of supervised and unsupervised learning algorithms that works on a consistent interface in Python. The library can also be used for data-mining and data analysis. The main machine learning functions that the Scikit-learn library can handle are classification, regression, clustering, dimensionality reduction, model selection, and preprocessing.

XGBoost:

XGBoost which tolerate for Extreme Gradient Boosting is one of the best Python packages for performing Boosting Machine Learning. Libraries such as LightGBM and Cat Boost are also fairly equipped with well-defined functions and methods. This library is built primarily for the purpose of implementing gradient boosting machines which are used to improve the performance and accuracy of Machine Learning Models.

Step by Step Guide to Create A Python Libraries:



MODULES DESCRIPTION:

Cyber Analysis:

Cyber threat analysis is a process in which the knowledge of internal and external information vulnerabilities pertinent to a particular organization is matched against real-world cyber-attacks. With respect to cyber security, this threat-oriented approach to combating cyber-attacks represents a smooth transition from a state of reactive security to a state of proactive one. Moreover, the desired result of a threat assessment is to give best practices on how to maximize the protective instruments with respect to availability, confidentiality and integrity, without turning back to usability and functionality conditions. A threat could be anything that leads to interruption, meddling or destruction of any valuable service or item existing in the firm's repertoire. Whether of "human" or "nonhuman" origin, the analysis must scrutinize each element that may bring about conceivable security risk.

Dataset Modification

If a dataset in your dashboard contains many dataset objects, you can hide specific dataset objects from display in the Datasets panel. For example, if you decide to import a large amount of data from a file, but do not remove every unwanted data column before importing the data into Web, you can hide the unwanted attributes and metrics, To hide dataset objects in the Datasets panel, To show hidden objects in the Datasets panel, To rename a dataset object, To create a metric based on an attribute, To create an attribute based on a metric, To define the geo role for an attribute, To create an attribute with additional time information, To replace a dataset object in the dashboard.

Data Reduction

Improve storage efficiency through data reduction techniques and capacity optimization using data reduplication, compression, snapshots and thin provisioning. Data reduction via simply deleting unwanted or unneeded data is the most effective way to reduce a storing's data

Risky User Detection

False alarm immunity to prevent customer embarrassment, High detection rate to protect all kinds of goods from theft, Wide-exit coverage offers greater flexibility for entrance/exit layouts, Wide range of attractive designs complement any store décor, Sophisticated digital controller technology for optimum system performance.

REALTED TO RSEARCH WORK:

Predicating User Personality by Mining Social Interactions in Face Book

The most frequently used procedure to obtain the user personality subsist of asking the user to fill in questionnaires. Though on one hand, it would be adorable to obtain the user personality as low-key. as feasible yet without agree the reliability of the model built. On the other hand, our assumption. Is that users with alike nature are await to show common behavioral patterns when interacting through virtual social networks, and that these patterns can be mined in order to predict the propensity of a user personality:

- Some personality feature can be predicted from user's interaction with Facebook.
- We present TP2010, a Facebook application to gather the users' personality.
- Different machine-learning techniques have been used to look for interaction patterns.
- The classifiers obtained have high accuracy when predicting user personality.
- The user number of friends and wall posts contribute to predict the user personality.

MachineLearning-BasedSentiment Analysis for Twitter Accounts

Despite the use of various machine-learning techniques and tools for sentiment analysis during elections, there is a dire need for a state-of-the-art approach. To deal with these challenges, the contribution of this paper includes the adoption of a hybrid approach that involves a sentiment analyzer that includes machine learning. Moreover, this paper also provides a comparison of techniques of sentiment analysis in the analysis of political views by applying supervised machine-learning algorithms such as Naïve Bayes and support vector machines (SVM).

Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning

Presents the imbalanced-learn API, a python toolbox to tackle the curse of imbalanced datasets in machine learning. The python package scikit-learn is used as a major support for the toolbox used. The two methods used here are as follows sampling: The process of reducing the number of samples is called sampling.

The implemented methods can be categorized into 2 groups:1. fixed under-sampling and 2. cleaning under-sampling Ensemble-learning: Ensemble methods offer an alternative to use most of the samples. This paper presented the foundations of the imbalanced-learn toolbox vision and API.

CONCLUSION

In this paper, we present a user-centric machine learning system which leverages big data of various security logs, alert information, and analyst insights to the identification of risky user. This system provides a complete framework and solution to risky user detection for enterprise security operation center. We describe briefly how to generate labels from SOC investigation notes, to correlate IP, host, and users to generate user-centric features, to select machine learning algorithms and evaluate performances, as well as how to such a machine learning system in SOC production environment. We also demonstrate that the learning system is able to learn more insights from the data with highly unbalanced and limited labels, even with simple machine learning algorithms. The average lift on top 20% predictions for multi neural network model is over 5 times better than current rule-based system. The whole machine learning system is implemented in production environment and fully automated from data acquisition, daily model refreshing, to real time scoring, which greatly improve and enhance enterprise risk detection and management. As to the future work, we will research other learning algorithms to further improve the detection accuracy.

REFERENCES

1) SANS Technology Institute. The 6 Categories of Critical Log Information 2013.

2)A. L. Buczak and E. Guven.A survey of data mining and machine learning methods for cyber security intrusion detection.

3)S. Choudhury and A. Bhowal. Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection, Smart Technologies and Managementfor Computing, Communication, Controls, Energy and Materials (ICSTM), 2015.

4) N. Chand et al. A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection, Advances in Computing, Communication, & Automation (ICACCA), 2016.

5) K. Goeschel. Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis, SoutheastCon, 2016.

6) M. J. Kang and J. W. Kang. A novel intrusion detection method using deep

neural network for in-vehicle network security, Vehicular Technology Conference, 2016.

ABOUT AUTHORS:

[1]**Mr.E.Nagalakshmi Reddy** is currently pursuing MCA in Siddharth Institute of Engineering & Technology, Puttur, Andhra Pradesh, India.

[2]**Mr. J.S. Ananda Kumar,**Assistant Professor in Dept. of MCA, Siddharth Institute of Engineering & Technology, Puttur, Andhra Pradesh, India.