Relatório do Dataset - Darknet

Marcus A. F. Dudeque[1], **Nicolas D. De Marco**[1]

[1]Departamento de Informática – Universidade Federal do Paraná (UFPR)

mafd17@inf.ufpr.br, ndm17@inf.ufpr.br

1. Problema a ser Resolvido

A *Darknet* é o espaço de endereço não usado da internet que não tem intenção de interagir com outros computadores. Devemos ficar alertas com qualquer comunicação da *Darknet* pelo fato de ela ter uma escuta passiva, ou seja, ele aceita pacotes de entrada, mas não pacotes de saída. Graças a ausência de hosts legítimos na darknet, qualquer tráfego é contemplado para não ser visto e é geralmente tratado como uma sonda ou uma configuração incorreta, além de serem conhecidos como buracos negros.

Muitos dos tráfego maligos utilizando a *Darknet* são possíveis com o auxílio da aplicação Tor ou VPN (*Virtual Private Network*), que ajudam a ocultar e redirecionar a origem de comunicações malignas. Com isso, queremos descobrir baseado nas informações fornecidas pelo dataset, como tipo de comunicação, tamanho de pacote, tempo de comunicação e outros, se a forma de tráfego é maligna, seja usando Tor ou VPN, para conseguir evitar e prevenir possíveis exploits classificados com essas características.

2. Visão Geral do Dataset

Para coleta das informações do Dataset que estamos usando, foram combinados dois datasets já existentes, o ISCXTor2016 e ISCXVPN2016 e seu respectivo tráfego Tor e VPN que correspondem às categorias do *Darknet*.

Para o dataset ISCXVPN2016 ser gerado, foi criada uma conta para ser possível utilizar serviços de chat e video conferência, capturando sessões regulares e sessões com VPN, tendo um total de 14 categorias de tráfego. O tráfego foi capturado utilizando *Wireshark* e tcpdump, gerando um total de 28GB de dados. Para a VPN foi usado um provedor de serviço de VPN externo conectando ele com o *OpenVPN* (modo UDP). Para gerar tráfego SFTP e FTPS também foi usando um serviço externo, além do Filezilla como cliente.

Para o dataset ISCXTor2016 ser gerado, foram criados três usuários para a coleta de tráfego no navegador e dois usuários para a parte de comunicação, como chat, FTP, p2p. Para a parte de tráfego não-Tor foi usada a parte de tráfego inofensivo do dataset de VPN citado anteriormente, já para a parte do tráfego Tor foram usadas 7 categorias de tráfego. O tráfego foi capturado novamente usando *Wireshark* e tcpdump, gerando um total de 22GB de dados.

Para facilitar o processo de classificação e marcação, foi capturado simultaneamente o tráfego de saída da estação de trabalho e o *gateway* do Tor, coletando um conjundo de pares de arquivos .pcap, um para o arquivo de tráfego regular(estação de trabalho) e um para o de tráfego Tor. Posteriormente, foram classificados os tráfegos capturados em dois passos, primeiro processando os arquivos .pcap gerados na estação de

trabalho, extraindo os fluxos e confirmando que a maioria deles foram gerados por uma aplicação X (Skype, ftps, etc.), depois foram marcados todos os fluxos do arquivo Tor .pcap como X.

No dataset que utilizamos, é usada uma abordagem de duas camadas, gerando tráfego *Darknet* e inofensivo na primeira camada, sendo o tráfego *Darknet* constituído por Stream de áudio, navegação, Chat, Email, P2P, transferência, Stream de Vídeo e VOIP que é gerado na segunda camada.

Praticamente todos os atributos apresentados no dataset são númericos, tendo por exemplo, IP, Porta, Protocolo utilizado, duração do Fluxo, quantidade de pacotes transitados, tamanho desses pacotes, tempos de espera, etc. Além disso, temos o atributo categórico Label que nos informa qual a categoria do tráfego feito. A característica extraída desse atributo categórico utilizando o método de OrdinalEncoder foi ['Audio-Streaming', 'Browsing', 'Chat', 'Email', 'File-Transfer', 'P2P', 'Video-Streaming', 'VOIP'] = [0, 1, 2, 3, 4, 5, 6, 7]

As classes do Dataset são compostas por tráfegos maliciosos, ou seja, compostos por acessos atráves de Tor ou VPN, e inofensivos, compostos por acessos ordinários. O total de amostras que temos no dataset é de 141.530, a distribuição das classes é a seguinte, 117.219 amostras inofensivas (82,82%) e 24.311 amostras malignas (17,18%), sendo 22.919 de VPN (16,19%) e 1.392 de Tor (0,99%).

3. Planejamento Futuro

Abordaremos o dataset de forma a complementar o trabalho realizado pelos responsáveis pela coleta dos dados. Para isso utilizaremos classificação, objetivando identificar com precisão se um acesso é ou não maligno com base nos outros dados coletados.

A ideia principal é classificar o acesso entre maligno ou inofensivo, de forma binária, e para isso um dos algoritmos de treino que utilizaremos será o de Regressão Logística (*Logistic Regression*), um algoritmo muito bom especialmente para classificações binárias como a que pretendemos fazer. Certamente testaremos com alguns outros algorimos de *machine learning* como *Random Forest* e *K-nearest Neighbors* para efeito de comparação, mas ainda assim LR será o principal objeto de estudo.

Classificando o acesso, será possível identificar se um acesso é maligno ou não, podendo assim bloquear acessos com tendência maior de serem classificados como malignos, garantindo maior segurança ao sistema/rede.

3.1. Pipeline de Execução Planejado

- Importar dados do *dataset* para o *notebook*;
- Elencar atributos;
- Tratar colunas categóricas para levantamento de características;
- Balancear dados para evitar viéses;
- Separar conjuntos de treino e teste;
- Treinar modelo;
- Testar modelo com LR e comparar resultados;
- Testar modelo com outros algoritmos e comparar resultados.