Análise de tráfego de rede para a identificação de objetos maliciosos

Fabio Antero de Pulpa Melo Junior, Leonardo Henrique Melo, Lourenço Alves Pereira Junior, Mauri Aparecido de Oliveira e Rafael Cenato dos Santos Silva

O aumento de dispositivos conectados à internet e a constante evolução das técnicas utilizadas para ataques cibernéticos sempre impuseram um grande desafio aos sistemas de segurança da informação. Como resposta, pode-se perceber cada vez mais a adoção da inteligência artificial, com foco nos dois aspectos ora mencionados. Este trabalho visa contribuir com a evolução das linhas de defesa utilizando aprendizado de máquina. Com aprendizado supervisionado será possível treinar algoritmos para detectar atividades maliciosas na rede por meio de análise dos metadados de rede, presentes no protocolo de comunicação NetFlow.

Neste artigo foram analisados os desempenhos dos algoritmos de Redes Neurais(MLP), kNN, Decision Tree, Logistic Regression e Naive Bayes, os quais foram submetidos às fases de treinamento, validação e testes, utilizando dados dos conjuntos NF-UNSW-NB15-v2 e NF-ToN-IoT-v2.

I. INTRODUÇÃO

Intrusion Detection Systems (IDS) visam detectar ataques e proteger os perímetros de redes privadas presentes em variadas organizações. A segurança cibernética ganha ainda mais destaque com o aumento do número de dispositivos de Internet das Coisas (IOT) conectados à *internet*. Os sistemas IDS tradicionalmente utilizam assinaturas de ataque para detecção, entretanto estes sistemas não são efetivos contra ataques inéditos, chamados *zero-day*, ou mesmo novas variações de ataques conhecidos. Diante disso, pesquisas indicam que os métodos de prevenção de ataques baseados em anomalias constituem opção mais eficaz [Khraisat et al., 2019].

Utilizando técnicas de Aprendizado de Máquina, um subcampo da Inteligência Artificial, pode-se gerar algoritmos capazes de aprender padrões complexos de ataques cibernéticos. A partir destes algoritmos, um IDS poderá identificar um comportamento anômalo em seu tráfego normal e imediatamente realizar alguma ação de contenção ou notificação.

II. REFERENCIAL TEÓRICO

Uma das grandes preocupações atuais é a criação de IDS que consigam identificar agentes maliciosos na rede. Alguns autores têm descrito formas de impedir os ataques por meio de filtros e *firewalls*. Conjuntamente, a aplicação de aprendizado de máquina tem se demonstrado um auxílio importante para a detecção de dados mal intencionados [Ahmad et al., 2021]. O uso de cada uma destas abordagens depende da aplicação em questão. Modelos por assinatura, por sua vez, tendem a ser mais leves. Por outro lado, os modelos de aprendizado por detecção de anomalias tendem a se adaptar melhor a mudanças nas estruturas dos ataques.

A variação crescente e evolução das ferramentas de ataque torna-se um importante desafio. A necessidade de modelos que consigam acompanhar a volatilidade das ameaças foi apontada por [de Carvalho Bertoli et al., 2021]. Paralelamente, os conjuntos de dados disponíveis para este tipo de análise expressam dificuldade em acompanhar a evolução dos agentes maliciosos, e muitas vezes estão desatualizados, bem como muitos não possuem rótulos separando as classes de dados benignos de dados maliciosos. O problema apontado por [Sarhan et al., 2021], consiste no fato de muitos conjuntos de dados não compartilharem os mesmos atributos, o que representa um impeditivo para a criação de modelos que possam ser replicados em uma aplicação no mundo real.

[Sarhan et al., 2021] desenvolveram um conjunto de *datasets* que possuem os mesmos atributos e avalia o desempenho de detecção dos *datasets* levando em conta uma abordagem binária com uma classificação entre dados benignos e dados maliciosos. Além disso, os autores realizam uma análise considerando os mesmos tipos de ataques em uma abordagem multi-classe. Entretanto, o autor não realiza testes para validar se a padronização dos atributos é suficiente para gerar um modelo capaz de identificar mais de uma classe de ataque em um mesmo *dataset*.

III. MATERIAIS E MÉTODOS

O *dataset* escolhido para a realização de experimentos na identificação de fluxos de pacotes maliciosos foi o NF-UNSW-NB15-v2 [Sarhan et al., 2021]. O formato baseado em NetFlow do conjunto de dados UNSW-NB15 [Moustafa and Slay, 2015], denominado NF-UNSW-NB15, foi expandido com recursos adicionais do NetFlow e rotulado com suas respectivas categorias de ataque. O número total de fluxos de pacotes é de 2.390.275, dos quais 95.053 (3,98%) são amostras de ataque e 2.295.222 (96,02%) são exemplos benignos. As amostras maliciosas são compostas por ataques de Exploits, Generic, Fuzzers, Backdoor, Denial of Service (DoS), Reconnaissance, Shellcode, Worms e Analysis.

IV. PRÉ-PROCESSAMENTO DOS DADOS

Os atributos IPV4_SRC_ADDR, IPV4_DST_ADDR, L4_SRC_PORT e L4_DST_PORT que contêm informações relativas à identidade (endereço IP e porta) de dispositivos ou agentes de rede, foram removidos para evitar que o modelo vincule os pacotes maliciosos a dispositivos específicos. Foram removidos ainda o atributo L7_PROTO, por não ter sido encontrada informação satisfatória sobre o conteúdo do atributo no trabalho apresentado por [Moustafa and Slay, 2015].

Além destes, foram removidos atributos que apresentaram valor constante em todo o conjunto de dados (TCP_URGENT_POINTER). Ainda, foram removidos atributos com alta correlação, identificados a partir de uma matriz de correlação. Neste contexto, foram removidos os atributos MIN_TTL, MAX_TTL, ICMP_IPV4_TYPE e ICMP TYPE.

Por fim, os atributos TCP_FLAGS, CLIENT_TCP_FLAGS e SERVER_TCP_FLAGS foram convertidos em atributos lógicos representando cada uma das flags contidas nos valores numéricos apresentados. Consequentemente, estes atributos foram apagados mantendo-se apenas os atributos individualizados. Além disso, os atributos TCP_FLAGS_BIN, CLIENT_TCP_FLAGS_BIN e SERVER_TCP_FLAGS_BIN foram removidos por serem atributos intermediários para a obtenção das *flags* individuais.

V. EXPERIMENTOS

Os modelos de aprendizado foram obtidos por meio de um processo de validação cruzada com 10 pastas. Levando em consideração apenas exemplos de teste do conjunto de dados NF-UNSW-NB15 em cada uma das pastas foram utilizadas três pastas internas para a obtenção dos melhores parâmetros para os modelos. Durante a avaliação dos modelos foi utilizada uma porção de dados, que não participou do treinamento das pastas, para teste de cada uma das iterações. Essa porção de dados representou 10% dos dados selecionados para a validação cruzada e foi diferente para cada uma das pastas.

Ao término do processo de treinamento, os resultados obtidos para cada uma das pastas foram agrupados de modo a permitir uma visão geral sobre o conjunto de dados. A efetividade de cada uma das técnicas de aprendizado utilizadas foi constatada pela métrica de avaliação F1-Score. Assim, foi possível a realização de uma comparação direta entre as diferentes técnicas de aprendizado de máquina para a detecção de fluxos de pacotes maliciosos, como é apresentado na Tabela 1.

O processo de validação proposto utilizou-se de uma parcela do conjunto de dados ToN-IoT-v2, o qual possui ataques de Ransomware, CrossSite Scripting (XSS), Password, Distributed Denial of Service (DDoS), Injection e Man in the Middle (MITM). Estes ataques não foram utilizados para a criação dos modelos e são desconhecidos pelo processo de classificação. Ainda, ressalta-se que essa parcela de dados utilizados para testes contém o mesmo conjunto de atributos utilizado para treinamento.

Assim, as etapas de reescala, exclusão de atributos e a alteração de dimensionalidade foram executadas sobre este novo conjunto de dados, visando prepará-los para o formato esperado pelos modelos.

Tabela 1 - Resultados obtidos após a avaliação dos modelos no conjunto de dados.

Modelo de Aprendizado	Média *F1-Score*	Desvio Padrão
Árvore de Decisão	0.9963	0.0004
Regressão Logística	0.9773	0.0405
Naive Bayes	0.8692	0.0039
Perceptron Multi Camadas	0.9883	0.0022
K-Vizinhos Mais Próximos	0.9143	0.0051

Tabela 2 - Resultados obtidos após a avaliação dos modelos no conjunto de dados ToN-IoT v2.

DecisionTree	LogisticRegression	NaiveBayes	MLP	kNN
0.4613	0.52221	0.559146	0.369953	0.29694

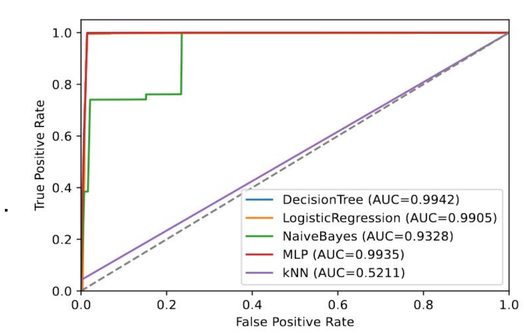

Figura 1 - Curva ROC dos resultados obtidos para um mesmo conjunto de dados.

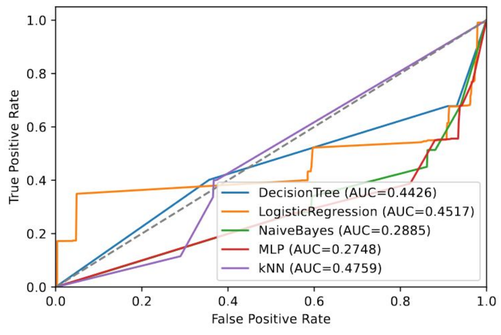

Figura 2 - Curva ROC dos resultados obtidos para a avaliação dos modelos no conjunto de dados ToN-IoT v2.

VI. CONCLUSÃO

Em virtude dos resultados obtidos na etapa de experimentos, pode-se observar que o modelo obteve bons resultados considerando o conjunto de ataques em que este foi treinado. Por outro lado, os resultados obtidos pela avaliação mostram que os modelos não foram capazes de identificar ataques desconhecidos, ou seja, para os quais nunca foram treinados, o que indica que esses modelos não terão um bom desempenho diante de ataques novos, conhecidos como *zero-day*. Levando em consideração a evolução constante das ferramentas de ataque utilizadas por agentes maliciosos, verifica-se a necessidade de um modelo capaz de identificar novos tipos de ataques.

Como trabalhos futuros sugere-se avaliar a capacidade da atualização de modelos através de outras técnicas de aprendizado de máquina, além de treinar os modelos utilizando uma maior variedade de conjunto de dados para melhorar a capacidade de generalização do modelo preditivo. Ainda é possível buscar maneiras para facilitar a rotulação de tráfego de rede em tempo real e gerar um esquema para a extração de atributos seguindo o protocolo Netflow.

REFERÊNCIAS

1. Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., and Ahmad, F. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1):e4150. 2021.
2. de Carvalho Bertoli, G., Júnior, L. A. P., Verri, F. A. N., dos Santos, A. L., and Saotome, O. Bridging the gap to real-world for network intrusion detection systems with data-centric approach. CoRR, abs/2110.13655. 2021.
3. Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. d. L. F. d. Inteligência artificial: uma abordagem de aprendizado de máquina. LTC. 2021.
4. Khraisat, A., Gondal, I., Vamplew, P., and Kamruzzaman, J. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity, 2(1):20. 2019.
5. Moustafa, N. and Slay, J. (2015). Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). Military Communications and Information Systems Conference (MilCIS), pages 1–6. 2015.
6. Sarhan, M., Layeghy, S., and Portmann, M. (2021). Towards a standard feature set for network intrusion detection system datasets. Disponível em: https://arxiv.org/abs/2101.11315. Acesso realizado em 26 de abril de 2022.
7. Systems, C. (2011). Cisco ios netflow version 9 flow-record format - white paper. Disponível em : https://www.cisco.com/en/US/technologies/tk648/tk362/technologies_white_paper09186a00800a3db9.pdf. Acesso realizado em 26 de abril de 2022.