Estudo da Influência das Features e do Desempenho dos Classificadores na Detecção de Anomalias em Redes

João Lucas Oliveira Mota[1], João Lucas Rodrigues da Silva[2], José Alberto Rodrigues Neto[2]

[1] Departamento de Engenharia de Teleinformática – Universidade Federal do Ceará (UFC)
Fortaleza – CE – Brasil

[2]Departamento de Computação – Universidade Federal do Ceará (UFC)
Fortaleza – CE – Brasil

{jlucasoliveira2002}@gmail.com, {joaolucas.rodrigues,jose.alberto}@alu.ufc.br

Abstract. *This document is a model and instructions for LaTeX. This and the sbc-template style define the components of your paper. Do not use symbols, special characters, footnotes, or math in the paper title or abstract.*

Resumo. *Este documento apresenta o modelo e o conjunto de instruções para artigos elaborados em LaTeX segundo o padrão da SBC. O arquivo sbc-template define os principais componentes do artigo, como título, autores e seções.*

1. Introdução

CHANGE ME Este artigo investiga a influência da seleção de atributos e do desempenho de diferentes classificadores na detecção de anomalias em tráfego de redes de computadores.

2. Dataset e pré-processamento

2.1. Dataset

O conjunto de dados utilizado foi obtido a partir da plataforma Kaggle, no repositório *Network Intrusion Detection*, disponibilizado publicamente por Sampada Bhosale [Bhosale 2018].

O dataset foi construído a partir da simulação de um ambiente de rede militar típico da Força Aérea dos Estados Unidos, desenvolvido para representar diferentes tipos de tráfego de rede, incluindo tanto comunicações legítimas quanto diversos tipos de ataques. Cada instância do conjunto de dados representa uma conexão de rede, contendo atributos como a quantidade de bytes transmitidos pela origem, a duração da conexão e outros parâmetros, conforme descrito na tabela apresentada no Apêndice.

Por fim, após as simulações realizadas de forma supervisionada, as conexões foram classificadas como tráfego normal ou como intrusão, fazendo com que a variável alvo seja do tipo binária, assumindo os valores Normal ou Anomalia.

2.2. Pré-processamento

CHANGE ME A aplicação de PCA como etapa de pré-processamento tem se mostrado eficaz para reduzir a dimensionalidade dos dados sem perda significativa de informação, preservando cerca de 99% da variância original mesmo com reduções superiores a 50% no número de atributos, conforme observado por Santos e Miani [Santos and Miani 2025].

3. Metodologia

3.1. Modelos Utilizados

3.1.1. K-Nearest Neighbors

Outro modelo que utilizamos foi o K-Nearest Neighbors (KNN). Este tipo de modelo é único porque, diferente dos outros que utilizam os dados de treinamento para aprender padrões e prever ataques, ele apenas armazena os dados de treinamento e compara a entrada com um número arbitrário (K) de casos semelhantes registrados.

3.1.2. Isolation Forest

O Isolation Forest (IF) é um algoritmo de aprendizado de máquina não supervisionado, projetado especificamente para a detecção de anomalias (ou outliers) em conjuntos de dados. A escolha do Isolation Forest foi motivada pelo fato de que, em alguns cenários, não é possível obter todos os dados rotulados necessários para a detecção de intrusões em redes, permitindo avaliar sua capacidade de identificar anomalias mesmo com ausência parcial de rótulos.

3.1.3. Random Forest

3.1.4. Multilayer Perceptron (MLP)

4. Resultados e Discussão

4.1. Resultados com K-Nearest Neighbors

Após alguns testes, os resultados mais precisos foram encontrados com $K = 1$, resultando em um F1-Score de 0.9946. Provavelmente, esse comportamento ocorre devido à grande quantidade de dados disponíveis para cada entrada de treinamento, tornando a maioria dos ataques evidentes. Valores maiores de K levaram ao sobreajuste e ao decréscimo progressivo do desempenho.

Após uma análise baseada na remoção individual dos atributos, observou-se que o parâmetro *hot* foi o mais decisivo para o modelo, reduzindo o F1-Score para 0.9931 quando removido. Além disso, a remoção de oito atributos resultou em um aumento do desempenho, atingindo um F1-Score de 0.9960, sendo o atributo *diff_srv_rate* o mais prejudicial, com impacto positivo de 0.0005 pontos após sua exclusão.

4.2. Resultados com Isolation Forest

O Isolation Forest apresentou desempenho inferior em relação aos demais modelos avaliados, obtendo acurácia de 0.748. Entretanto, a acurácia não é a métrica mais adequada para conjuntos de dados desbalanceados, como é comum em cenários de detecção de anomalias, pois pode ser influenciada pela predominância da classe majoritária (tráfego normal).

A Tabela 1 apresenta as métricas de desempenho do modelo.

Table 1. Métricas de desempenho do Isolation Forest

Métrica	Classe Normal	Classe Anomalia
Precision	1.00	0.65
Recall	0.53	1.00
F1-Score	0.69	0.79
Support	2690	2349

Observa-se que todos os eventos classificados como normais eram, de fato, normais. Entretanto, o modelo apresentou dificuldades em identificar corretamente eventos normais, obtendo recall de apenas 53%, o que indica a geração de um número significativo de falsos positivos. Apesar disso, o modelo não apresentou falsos negativos, classificando corretamente todos os ataques detectados como anomalias.

Esse comportamento pode gerar retrabalho para equipes de segurança, uma vez que eventos normais classificados como ataques exigem análise manual adicional.

Agradecimentos

Os autores agradecem à Universidade Federal do Ceará (UFC) pelo suporte institucional.

References

Bhosale, S. (2018). Network intrusion detection. Kaggle. Acesso em 2025.

Santos, K. C. and Miani, R. S. (2025). Impacto da redução de dimensão e seleção de atributos na generalização de modelos de detecção de intrusão. In *Anais do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, Uberlândia, MG, Brazil. SBC.