Seleção de Características com Alta Quantidade de Informação para Sistemas de Detecção de Intrusão baseada no Conjunto de Dominância de Pareto

Guilherme Nunes Nasseh Barbosa[1] **e Diogo Menezes Ferrazani Mattos**[1]

[1] LabGen/MídiaCom – TET/IC/PPGEET/UFF
Universidade Federal Fluminense (UFF)
Niterói, RJ – Brasil

{gnasseh,diogo_mattos}@id.uff.br

Abstract. *The COVID-19 pandemic has driven a change in the profile of Internet use, fostering an increase in attacks and new threats to institutions, which until then had been little targeted. In this new scenario, threat detection and prevention tools tend to be replaced by machine learning-based solutions that require efficient execution. This article proposes an efficient method for feature selection for machine learning using the Pareto frontier. The proposal minimizes the Pearson correlation and the Mutual Information between pairs of selected features. The selected dominant features were applied to three machine-learning models for classifying malicious streams. The proposed method was efficient compared to other methods, as it allows using fewer features to achieve similar accuracy, precision, and recall values, reducing training and validation time.*

Resumo. *A pandemia de COVID-19 impulsionou a mudança no perfil de uso da Internet, o que fomentou o aumento de ataques e novas ameaças a instituições, até então, pouco visadas. Nesse novo cenário, ferramentas de detecção e prevenção de ameaças tendem a ser substituídas por soluções baseadas em aprendizado de máquina, que exigem execução eficiente. Este artigo propõe um método eficiente para a seleção de características para o aprendizado de máquina, utilizando a fronteira de Pareto. A proposta minimiza a correlação de Pearson e a Informação Mútua entre pares de características selecionadas. As características dominantes selecionadas foram aplicadas a três modelos de aprendizado de máquinas para classificação de fluxos maliciosos. O método proposto apresentou eficiência quando comparado a outros métodos, pois permite utilizar menos características para atingir valores similares de acurácia, precisão e revocação, diminuindo o tempo de treinamento e validação.*

1. Introdução

O tráfego na Internet vem aumentando nos últimos anos devido a diversos fatores e, como consequência, o número de ataques virtuais também cresce proporcionalmente. Em relatório recente, a empresa Checkpoint aponta que no terceiro trimestre de 2022, os ataques virtuais cresceram 28% , comparado com o mesmo período de 2021[1]. O setor mais afetado neste período foi o de educação e pesquisas, o qual sofreu 18%

[1]Disponível em https://blog.checkpoint.com/2022/10/26/third-quarter-of-2022-reveals-increase-in-cyberattacks/.

de aumento em número de ataques em comparação ao mesmo período do ano anterior, enquanto instituições de saúde foram os principais alvos de *ransomware*, aumento de 5% em relação a 2021. O aumento de ataques nestes dois segmentos é creditado a reflexos da pandemia de COVID-19. Instituições de pesquisas tornaram-se alvos de ataques cibernéticos no momento em que buscavam respostas para esta doença infecciosa. O foco em instituições de saúde provém da complexidade existente na operação cotidiana, uma vez que, com a digitalização de informações de prontuários, a criptografia destas informações pode inviabilizar por completo o funcionamento de um hospital. A utilização de dispositivos pessoais também contribui para a disseminação de ataques por não possuírem as mesmas políticas de segurança adotadas em empresas. Com o alto volume de tráfego gerado, sistemas clássicos de detecção de anomalias baseados em assinatura, como Sistemas de Detecção de Intrusão (*Intrusion Detection System* – IDS) e Sistemas de Prevenção de Intrusão (*Intrusion Prevention System* – IPS), têm se mostrado poucos eficazes. Estes sistemas implicam um grande um intervalo entre a criação da assinatura e sua implementação [Matin e Rahardjo, 2019]. Por sua vez, sistemas baseados em anomalias requerem tempo de aprendizado para estabelecer o perfil de uso normal da rede, contudo, tal perfil está atualmente em mudança devido à introdução constante de novas aplicações. Ferramentas baseadas em aprendizado de máquinas são soluções promissoras para análises de grandes fluxos de dados com alta complexidade [Medeiros et al., 2019].

Aplicações de detecção de ameaças baseadas em aprendizado de máquinas possuem maior complexidade, quando comparadas a IDS por assinatura, pois exigem a análise de grandes volumes de dados em um pequeno intervalo de tempo com processamento em fluxo. Ameaças podem ser ofuscadas em grandes volumes de dados, o que torna sua detecção complexa. Os modelos de aprendizado de máquinas precisam ser ajustados continuamente em função dessa ofuscação. O tempo de treinamento e refinamento desses modelos torna-se um desafio para a classificação de ameaças em tempo real. A seleção de características em conjuntos de dados também possui particularidades. As características de treinamento devem ser ajustadas de maneira contínua para tornar os modelos eficientes em cada ambiente monitorado. Dessa forma, a classificação de tráfego anormal em redes de computadores baseado em aprendizado de máquina deve considerar o menor tempo de treinamento possível para diminuir o tempo de resposta a incidentes. A análise quantitativa e qualitativa das características está diretamente relacionada à eficiência dos modelos treinados, pois quanto maior o número de características, maior é o tempo de treinamento dos modelos, porém, não necessariamente mais preciso é o modelo [Andreoni Lopez et al., 2018].

Este artigo propõe um método eficiente de seleção de características baseado na otimização multiobjetivo da seleção de características com mínima correlação linear e mínima informação mútua para treinamento de modelos de classificação aplicados em fluxos de redes. A ideia chave da proposta é selecionar o conjunto de características que compartilhem a menor quantidade de informação entre si, para, assim, obter um conjunto final em que haja o mínimo de características redundantes entre si. A proposta consiste na utilização do menor número possível de características para treinamento dos modelos, de modo que seja possível uma redução no tempo de classificação de anomalias. Para isso, as características do fluxo são avaliadas através da Correlação de Pearson e Informação Mútua para minimização multiobjetivo de tais funções. A minimização multiobjetivos é realizada através do conjunto de dominância de Pareto. A proposta é implementada e

avaliada sobre um conjunto de dados real que contém tráfego normal e ameaças coletados de 373 usuários de banda larga doméstica na cidade do Rio de Janeiro [Lopez et al., 2017].

Trabalhos anteriores incorporam o aprendizado de máquina no monitoramento de ambientes através de detecção de intrusão de rede (IDS) [Di Mauro et al., 2021]. Logo, é possível estabelecer perímetros em ambientes micro-segmentados, evitando ataques laterais em recursos compartilhados [Arifeen et al., 2021]. O principal desafio da seleção de características para classificadores, recai sob a qualidade de informação que cada característica agrega ao modelo. O método Chi-Quadrado pode ser útil ao avaliar a independência de características categóricas, contudo, necessita de processamento para normalização dos dados [Thakkar e Lohiya, 2021]. Os modelos de classificação binária são utilizados amplamente no monitoramento de rede, porém, a extração de informações das características é essencial para a precisão [Kasongo e Sun, 2019]. Diversos classificadores carecem de informações precisas para o treinamento e detecção de ameaças em função do conjunto de dados utilizado, pois tais conjuntos possuem grande enviesamento ao analisar a correlação das características [Silva et al., 2022]. Diferentemente de trabalhos anteriores, a proposta foca na seleção de característica com processamento reduzido ao passo que busca um conjunto de dados final com baixo enviesamento.

O restante do artigo está organizado da seguinte forma: A Seção 2 elenca os trabalhos relacionados. O problema da seleção do conjunto ótimo de características é analisado na Seção 3. A seleção de características multiobjetivos é proposta na Seção 4. A Seção 5 avalia a proposta e discute os resultados obtidos. A Seção 6 conclui o trabalho.

2. Trabalhos Relacionados

A detecção de ameaças em tráfego de rede objetiva identificar antecipadamente padrões que possam comprometer a integridade e a disponibilidade das redes de computadores. Dessa forma, a seleção de características para treinamento de modelos de aprendizado de máquinas deve ser amplamente avaliada como uma importante etapa de pré-processamento dos dados. A seleção de características busca aquelas que possuam alta quantidade de informação e tornem os modelos mais precisos e acurados. Lopez *et al.* propõem um modelo de pré-processamento não supervisionado para correlação e normalização de características em fluxo de redes [Andreoni Lopez et al., 2019]. Os autores comparam o modelo proposto com outros algoritmos clássicos, tais como a análise de componentes principais (*Principal Component Analysis* - PCA) e a seleção sequencial de características (*Sequential Feature Selection* - SFS). A proposta consiste na implementação de dois algoritmos. O primeiro para normalização das características do conjunto de dados, no qual os dados são submetidos a uma distribuição normal para mudança de escala dos valores, para que fiquem compreendidos em um intervalo entre -1 e 1. O segundo algoritmo executa a seleção de característica de forma proporcional à variância das características e inversamente proporcional à correlação entre características. O trabalho foca no processamento em fluxo dos dados.

Ma *et al.* analisam as requisições HTTP para identificar tráfego de rede suspeitos. Os localizadores uniformes de recursos (*Uniform Resource Locator* - URLs) das requisições são transformadas em vetores de peso baseados em critérios estatísticos e posteriormente são convertidos em vetores utilizando a técnica *k-gram* para processamento de linguagem natural. Os resultados da transformação são utilizados como entrada no

classificador SVM[Ma et al., 2021]. Kim *et al.* [Kim e Cho, 2018] propõem um modelo híbrido de rede neural utilizando Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNN) e redes neurais de memória de curto e longo prazo (*Long Short-Term Memory* - LSTM), para extração automática de características a partir do tráfego web. A proposta baseia-se na análise espaço-temporal do tráfego HTTP que contenha determinados padrões e reduz o espectro de informações relacionado ao contexto temporal utilizando uma única camada de uma rede neural convolucional. A saída dessa rede é utilizada como entrada na rede LSTM para reduzir as variações temporais. Por fim, a saída da rede LSTM é conectada a uma Rede Neural Profunda para realizar a classificação. Embora os trabalhos visem a realização de engenharia de características, a redução da dimensionalidade não é o foco.

Wang *et al.* propõem uma arquitetura distribuída para detecção de anomalias em redes virtualizadas para nós e enlaces físicos. Os autores aplicam o algoritmo de aprendizado de máquina de vetor de suporte de classe única (*One-Class Support Vector Machines* - 1-SVM) para detectar anomalia em cada nó do ambiente de virtualização e a análise de correlação canônica (CCA)[Wang et al., 2022]. O algoritmo 1-SVM identifica como anomalia qualquer ponto que esteja fora da superfície de decisão definida pelos vetores de suporte. Diferentemente de abordagens que consideram o problema de identificação de ameaças como um problema de classificação de amostras, o trabalho de Wang *et al.* mapeia a identificação de anomalias sobre o problema de identificação de pontos discrepantes abordado pelo algoritmo 1-SVM.

Farrugia *et al.* objetivam determinar quais características possuem um impacto relevante nos modelos de detecção de anomalias para detecção de fraudes em blockchain utilizando a rede Ethereum. Após determinar as principais características, os autores utilizaram o modelo de classificação XGBoost para determinar contas maliciosas envolvidas em transações [Farrugia et al., 2020]. Garg *et al* propõem um método para detecção de anomalias em redes utilizando *Grey Wolf Optimizer* (GWO) e redes neurais convolucionais. Na proposta os autores utilizam o algoritmo GWO para seleção multiobjetivo das características e a Rede Neural Convolucional para classificação das anomalias [Garg et al., 2019].

A proposta deste artigo diferencia-se de trabalhos anteriores por considerar a seleção de características um problema de otimização multiobjetivo. Este artigo considera a hipótese de que as melhores características a serem selecionadas para o treinamento dos modelos de aprendizado de máquina são aquelas que apresentam a menor correlação linear e, concomitantemente, a menor informação mútua. A hipótese adotada neste artigo baseia-se no fato de que características altamente correlacionadas agregam informação redundante ao modelo treinado [Andreoni Lopez et al., 2019]. Contudo, argumenta-se que a correlação de Pearson por si só não captura relações não-lineares entre as características e, portanto, assume-se a minimização conjunta com a informação mútua, já que essa métrica é capaz de traduzir relações não-lineares entre variáveis.

3. Desafios para a Seleção Ótima de Características

A seleção de características em problemas de aprendizado de máquina consiste no processo de selecionar um subconjunto de características relevantes para uso na construção do modelo. O objetivo da seleção de características é identificar um pequeno

número de características que são as mais relevantes para o problema, de modo a descartar as características irrelevantes ou parcialmente relevantes. A seleção de características é importante por ajudar a reduzir a complexidade do modelo e melhorar a sua interpretabilidade, por melhorar a generalização do modelo, reduzindo o sobreajuste e, por consequência, acelerar o treinamento e o teste do modelo [Andreoni Lopez et al., 2019]. As abordagens diferentes para a seleção de características classificam-se em:

- **métodos de filtragem** que selecionam características com base em testes estatísticos ou heurísticas;
- **métodos de envelopamento** *wrapper* que empregam um algoritmo de aprendizado para avaliar o desempenho de diferentes subconjuntos de características e, então, escolher aquele que apresenta o melhor desempenho;
- **métodos incorporados** que selecionam as características como parte do processo de treinamento do algoritmo de aprendizado;
- **métodos híbridos que combinam abordagens de métodos anteriores.**

A escolha adequada do subconjunto de características para uso em um modelo de aprendizado de máquina, no entanto, está sujeita a diversos desafios. À medida que o número de características de um problema aumenta, a quantidade de dados necessária para estimar com precisão as relações entre as características selecionadas e a variável alvo também aumenta. Isso pode levar ao sobreajuste e à degradação da generalização do modelo para novos dados. Tal desafio é conhecido como a maldição da dimensionalidade. Ademais, características altamente correlacionadas introduzem informações redundantes e não agregam valor ao modelo. Identificar e selecionar um subconjunto de características não correlacionadas é um desafio. Paralelamente, identificar e remover características irrelevantes também é uma tarefa árdua, especialmente se a característica estiver correlacionada com a variável alvo [Silva et al., 2022], o que pode introduzir um viés no modelo final. A análise combinatória de subconjunto de características muita das vezes não é uma solução viável, pois há a explosão do espaço de características. Caso o número de características em um conjunto de dados seja grande, o número de subconjuntos de características possíveis pode ser extremamente grande, tornando computacionalmente inviável avaliar todos eles. Por fim, a relação de compromisso entre o desempenho do modelo e a sua interpretabilidade também são um ponto de atenção. Selecionar um pequeno número de características tende a melhorar a interpretabilidade do modelo, mas também pode resultar em perda de desempenho global do modelo.

4. Seleção de Característica Multiobjetivos Proposta

A proposta deste trabalho consiste em um método de filtragem para seleção de características de um fluxo de rede para treinamento de classificadores. O método é uma heurística que se vale de métricas estatísticas para filtrar as características que são consideradas ótimas. A proposta baseia-se na otimização multiobjetivos para a minimização da correlação e da informação mútua entre as características selecionadas. A fim de reduzir o custo computacional para o cálculo das características que minimizam a correlação e a informação mútua entre as características selecionadas, a solução proposta para o problema de otimização é encontrar a fronteira de Pareto para o problema proposto. Assim, são selecionadas as características que estão na fronteira de Pareto. Para tanto, são calculadas a Correlação de Pearson e a Informação Mútua para pares de características, para

avaliar o grau de dominância que um par de características possui em relação aos demais. Ressalta-se que a complexidade do método é amortizada através do cálculo *a priori* das matrizes de correlação linear e de informação mútua entre todas as características. Desse modo, a avaliação de qualquer par de características reside na consulta às matrizes de correlação e de informação mútua. O método proposto de seleção de características divide-se em quatro etapas e é denominado Método Variável de Seleção de características, mostrado na Figura 1.

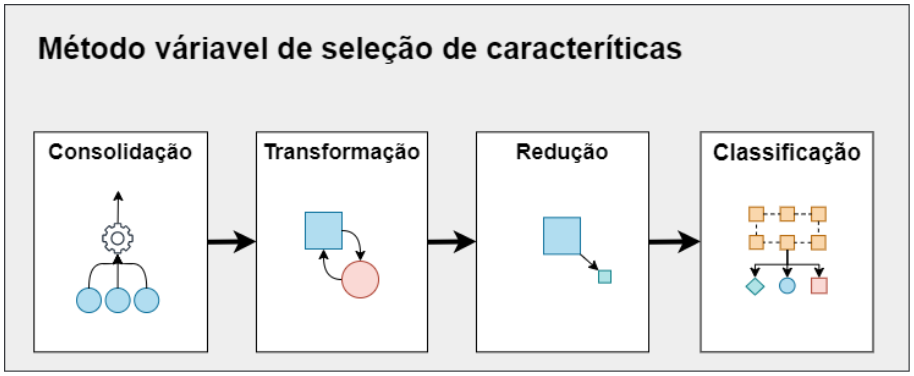


Figura 1. O método proposto é composto por quatro etapas: consolidação, transformação, redução e classificação. A seleção de características reside nas etapas de consolidação, transformação e redução. A etapa de classificação foca no uso do modelo reduzido para a classificação de tráfego.

4.1. Consolidação dos dados

A primeira etapa da proposta realiza o pré-processamento do conjunto de dados. Nesta etapa, são removidas as características categóricas e que identificam um fluxo, tais como a 5-tupla formada por IP de origem e destino, portas de origem e destino e protocolo de transporte. Embora tais características carreguem informações importantes para a identificação do fluxo, seu uso em modelos de classificação tendem a contribuir para o enviesamento do modelo e consequente especialização exacerbada do modelo, levando ao sobreajuste. Em modelo geral, essas características têm baixa relevância para a detecção ou classificação de ameaças por serem singulares, ou seja, para cada fluxo. As características ou a combinação de seus valores podem ser detectados como *outliers* pelos classificadores, não contribuindo de modo significativo para a identificação de ameaças. A proposta da consolidação é reduzir a dimensionalidade dos dados para otimizar a execução das etapas seguintes, evitando o consumo de tempo de computação com características que não agregam qualidade ao modelo final de classificação.

4.2. Transformação

A transformação dos dados é a parte de maior complexidade, pois a escolha de características sem relevância diminui o desempenho do treinamento e, consequentemente, da classificação de ameaças [Abdollahzadeh e Gharehchopogh, 2022]. Inicialmente, é calculada a matriz de Correlação de Pearson entre as características do conjunto de dados consideradas válidas na etapa de consolidação. Esta é uma ferramenta útil para avaliar a correlação entre duas variáveis contínuas, no qual uma matriz quadrática é construída. Os resultados da correlação de Pearson são valores compreendidos entre $-1 \leq \rho \leq 1$, no

qual 1 representa as características que possuem relação linear direta, enquanto valores negativos até -1 representam características com correlação linear inversa. O coeficiente ρ é calculado através da média μ e desvio padrão σ sendo descrito através de

$$\rho(A, B) = \frac{1}{N-1} \sum (\frac{A_i - \mu_A}{\sigma_A})(\frac{B_i - \mu_B}{\sigma_B}). \quad (1)$$

Em seguida, é calculada a Informação Mútua para cada par de características. A informação mútua representa o valor da entropia condicional de duas variáveis aleatórias. A entropia da informação, também conhecida como Entropia de Shannon, é uma medida para analisar o grau de incerteza ou concentração da distribuição da informação. A Informação Mútua representa o grau de influência em relação à classe alvo do conjunto de dados[Thakkar e Lohiya, 2021]. A Informação Mútua entre X e Y é utilizada para medir a quantidade de informações compartilhadas por X e Y [Li et al., 2017] é expressa por

$$IG(X, Y) = H(X) - H(X|Y), \quad (2)$$

em que H(X) é a entropia da classe alvo (X) descrita por

$$H(X) = -\sum_{i=1}^{n} (p_i) \log(p_i), \quad (3)$$

e H(X|Y) é a entropia condicional de X dada uma variável discreta aleatória Y e determinada por

$$H(X|Y) = -\sum_{y_j \in Y} P(y_j) \sum_{x_i \in X} P(x_i|y_j) log(P(x_i|y_j)). \quad (4)$$

No caso de seleção de características, propostas anteriores consideram individualmente a correlação linear ou a informação mútua como critérios de seleção das características no modelo final. O método ora proposto, no entanto, considera ambas as métricas simultaneamente.

4.3. Redução

Após a etapa de transformação das características, é aplicado o conceito de eficiência de Pareto para determinar um conjunto de soluções ótimas, conhecido com Fronteira de Pareto. Diversos problemas podem ser formulados para atingir objetivos singulares ou múltiplos. No entanto, esses objetivos podem conflitar entre si, dificultando a escolha de uma solução ótima. Em casos nos quais duas ou mais soluções possíveis devem ser comparadas, pode-se utilizar o conceito da Fronteira de Pareto [Viduto et al., 2012]. Tal conceito baseia-se na ideia de que dois objetivos possíveis para um único problema no qual existem diversas opções. Os pontos de um plano cartesiano que possuem as melhores condições são chamados de solução ótima ou pontos dominantes. A otimização de Pareto não define explicitamente qual solução deve ser obrigatoriamente escolhida, apenas restringe as opções. É possível formar uma área contendo todas as opções possíveis e, nesta área, existe uma borda definida pela Fronteira de Pareto, na qual nenhum ponto a ultrapassa. Os pontos que se encontram na fronteira possuem maior dominância sobre

os demais. Dessa forma, utilizando pares de características obtidos através da fronteira de Pareto que minimizam a correlação de Pearson e a Informação Mútua, é possível reduzir as características de acordo com as propriedades de uma janela de tempo ou de um fluxo de dados específico sem perder a generalização na classificação. A ideia principal é selecionar as características que se encontram na fronteira para treinar os modelos sem que seja necessário utilizar todas as características do conjunto de dados, pois, por mais eficiente que seja a utilização de mais características, essa condição onera o tempo de treinamento do modelo.

4.4. Classificação

A etapa de classificação tem como objetivo determinar se um fluxo de dados possui comportamento anormal. Para isso, diversos modelos são utilizados nessa etapa, todos eles com características distintas. Foram utilizados três modelos de aprendizado de máquina para classificação de tráfego malicioso. Para validar a escolha das características pela abordagem proposta, foram utilizados os classificadores Árvore de Decisão, Árvore Aleatória e *XGBoost*. Esses modelos são amplamente utilizados para a classificação de anomalias em diversos cenários [Silva et al., 2022]. Um protótipo da proposta foi implementado e os algoritmos de classificação utilizados foram implementados através das bibliotecas Scikit-Learn e XGBoost, ambas para Python. O modelo **Árvore de Decisão** é baseado no aprendizado supervisionado no qual seu objetivo principal é determinar valores discretos de uma variável alvo, aprendendo regras simples de decisão extraídas a partir da análise de características previamente definidas. O modelo **Árvore Aleatória** é um meta estimador que ajusta diversas classificações de uma árvore de decisão em subamostras do conjunto de dados. Esse modelo utiliza a média para melhorar a precisão preditiva e controlar o sobreajuste. O **XGBoost** é um algoritmo de aprendizado de máquina baseado em um conjunto de árvore de decisão que utiliza a abordagem de reforço do gradiente para a otimização da acurácia final [Farrugia et al., 2020].

5. Avaliação e Resultados

A proposta é avaliada sobre um conjunto de dados contendo fluxos de uma rede real de acesso banda larga doméstica da cidade do Rio de Janeiro. Para tanto, foi implementado um protótipo do método de seleção de características proposto. O treinamento e a validação da proposta foram realizados utilizando algoritmos de aprendizado de máquinas para um conjunto de dados rotulado de uma grande operadora de telecomunicações do Brasil [Lopez et al., 2017]. O conjunto de dados utilizado contém o tráfego de 373 usuários de banda larga residencial coletados na cidade do Rio de Janeiro. A coleta realizada utilizou o protocolo NetFlow e ocorreu durante uma semana, entre os dias 24 de fevereiro a 4 de março de 2017, tendo 46 características, além do rótulo que define um fluxo legítimo ou uma ameaça. As ameaças são determinadas a partir de alarmes de um sistema de detecção de intrusão (IDS). A rotulagem provida pelo IDS é a verdade básica (*ground truth*) do conjunto de dados. Para realizar o pré-processamento, treinamento e validação da proposta, o ambiente utilizado é composto por uma máquina com processador Intel Core i7 9700 com frequência de 3,00 Ghz, 16GB de RAM DDR4 e placa de vídeo NVIDIA GTX 1660 SUPER com 6GB de RAM GDDR5. O sistema operacional utilizado é o Ubuntu 18.04. A linguagem de programação utilizada foi Python

3.6 [2] com a biblioteca Sklearn[3].

Foram avaliados quatro cenários para seleção de características. O primeiro compreende a utilização de todas as características do conjunto de dados, totalizando 41 características[4]. O segundo cenário baseia-se na seleção que utiliza a Fronteira de Pareto para determinar os pares de características que minimizam a informação mútua e a correlação linear entre si. O terceiro cenário refere-se a uma seleção dos pares de características que maximizam simultaneamente a Correlação de Pearson e a Informação Mútua entre si. O quarto cenário considera uma seleção das características mais centrais do Cenário 3 e, portanto, evita a seleção de pares de características com alta correlação linear ou alta informação mútua para o modelo final, selecionando somente as características que possuem informações relevantes, eliminando as características com informações similares. Para os três últimos cenários, foram utilizadas 29 características do conjunto de dados. Na etapa de consolidação de dados, foram removidas 17 características do conjunto de dados, sendo as categóricas (*timestamp, srcip, dstip, srcport, proto, dstport*) que identificam o fluxo, as que possuem valores nulos em todos os fluxos (*td-active, min-idle, mean-idle, max-idle, std-idle, furg-cnt, burg-cnt*) e características com valores idênticos (*duration, mean-active, max-active, min-active*).

Inicialmente, foram calculadas as matrizes de Correlação de Pearson para dois dias de análise do conjunto de dados. A Figura 2 apresenta o mapa de calor para facilitar a visualização das correlações [Silva et al., 2022]. Cada índice dessas matrizes representa uma correlação com valores entre -1 e 1. A diagonal principal representa a correlação com a própria característica, tendo sempre valor igual a 1. É possível notar que cada dia possui correlações distintas. Isso pode indicar uma anomalia específica ou um conjunto delas, ou simplesmente indicar variações perenes dos fluxos, visto que, mesmo que legítimos, estão sujeitos à sazonalidade. A escala de cor indica que quanto mais próximo de 1 maior a correlação entre as características, enquanto valores negativos indicam que as características são inversamente relacionadas. Em seguida, foi realizado o cálculo da Informação Mútua. Essa métrica é considerada como a redução da incerteza de uma variável aleatória dado o conhecimento da outra. A métrica foi normalizada para valores entre 0 e 1. Após o cálculo da Correlação de Pearson e da Informação Mútua, os valores foram utilizados para determinar a Fronteira de Pareto para os cenários 2, 3 e 4.

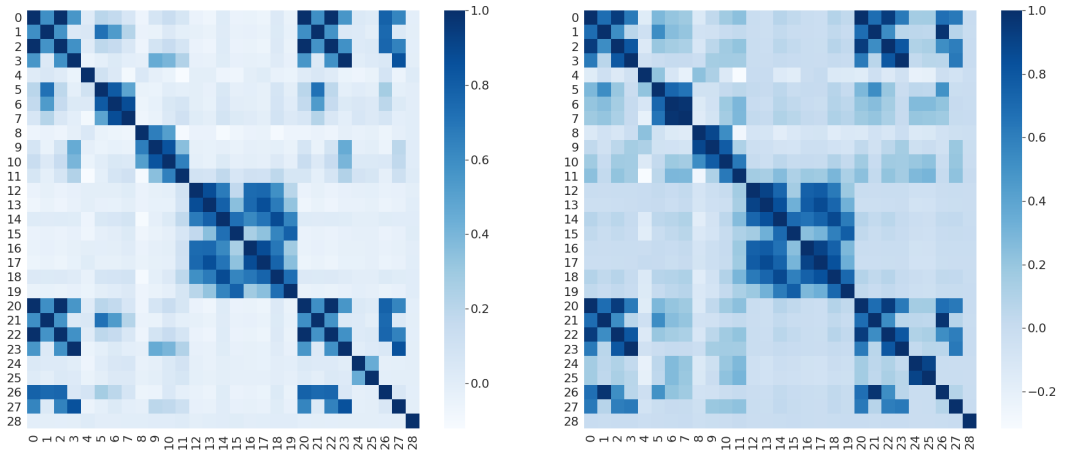
O primeiro cenário avaliado utiliza todas as características do conjunto de dados. Nesse cenário, embora as métricas de desempenho dos modelos tenham sido satisfatórias, o tempo de treinamento e validação foi superior a todos os outros cenários. Isso indica que, para treinamento de classificadores de tráfego, ao se utilizar todas as características disponíveis para o fluxo, o custo para treinamento é consideravelmente alto, independente do classificador utilizado.

O Cenário 2 compreende as características selecionadas pela Fronteira de Pareto para a minimização da informação mútua e da correlação linear. A Figura 3 apresenta a Fronteira e os pares de características selecionadas em azul, sendo elas *dscp, mean-fiat, mean-fpktl, min-biat, min-fiat, min-fpktl, std-biat*. Essas 7 características foram utilizadas

[2]Disponível em https://www.python.org/.

[3]Disponível em https://scikit-learn.org/stable/.

[4]O conjunto de dados apresenta 46 características totais, porém a 5-tupla que caracteriza cada fluxo foi desconsiderada.



(a) Correlação de Pearson para o dia 26/02/2017. (b) Correlação de Pearson para o dia 03/03/2017.

Figura 2. Mapas de calor da correlação de Pearson de 29 características para dois dias analisados. a) Correlação do dia 01/03/2017 e b) correlação do dia 03/03/2017.

para treinamento e validação dos modelos. As características representam, respectivamente o tempo médio entre pacotes transmitidos para frente, o tempo mínimo entre transmitidos no sentido contrário, o tempo mínimo entre pacotes pacotes transmitidos para frente, o tamanho do menor pacote transmitido enviado para frente e o desvio padrão no tempo médio entre pacotes transmitidos no sentido contrário.

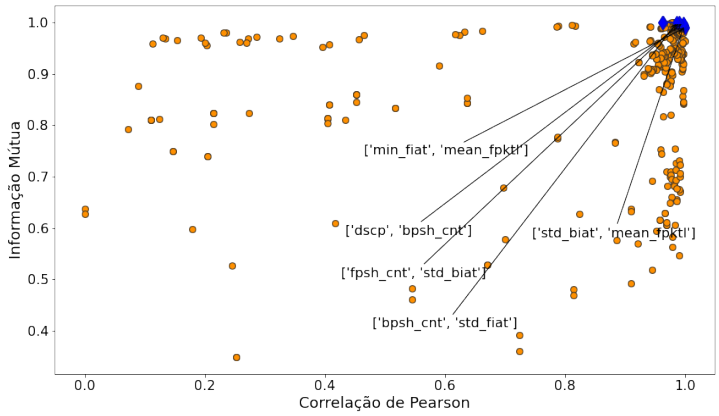
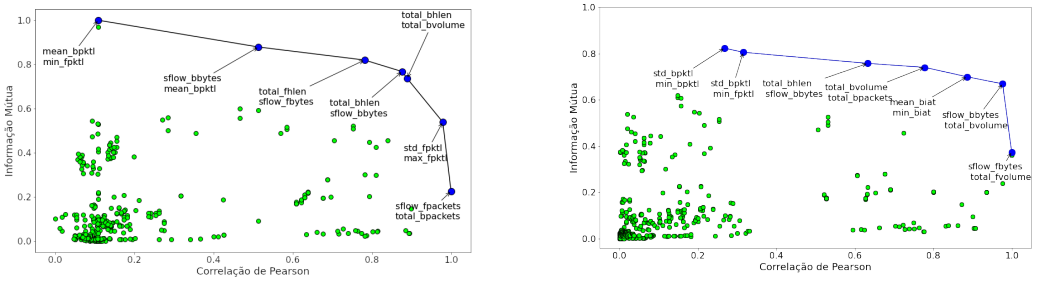


Figura 3. Características escolhidas pela Fronteira de Pareto com maior Correlação de Pearson e maior Informação Mútua.

A Figura 4 apresenta os pares de características referente ao Cenário 3, no qual há a maximização da informação mútua e da correlação entre características selecionadas. É possível notar que para cada dia analisado, os pares de características são distintos. No entanto algumas características se mostraram recorrentes, podendo indicar que possuem forte dominância sobre as demais em um fluxo de rede. As características *sflow-fbytes, sflow-bbytes, total-bvolume, total-bpackets, total-bhlen, min-fpktl* foram as recorrentes

para ambos os dias. As características utilizadas nesse cenário são respectivamente o número médio de bytes transmitidos em um sub-fluxo na direção para frente, o número médio de bytes transmitidos em um sub-fluxo na direção oposta, o volume total em bytes do pacotes transmitidos para frente, o número de pacotes na direção oposta, a quantidade total de bytes utilizados por cabeçalhos na direção oposta e o tamanho do menor pacote transmitido enviado para frente.

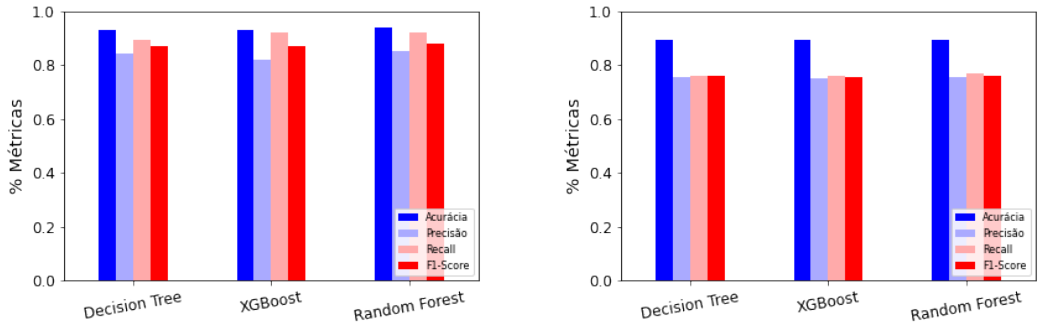


(a) Fronteira de Pareto dia 26/02/2017. (b) Fronteira de Pareto dia 03/03/2017.

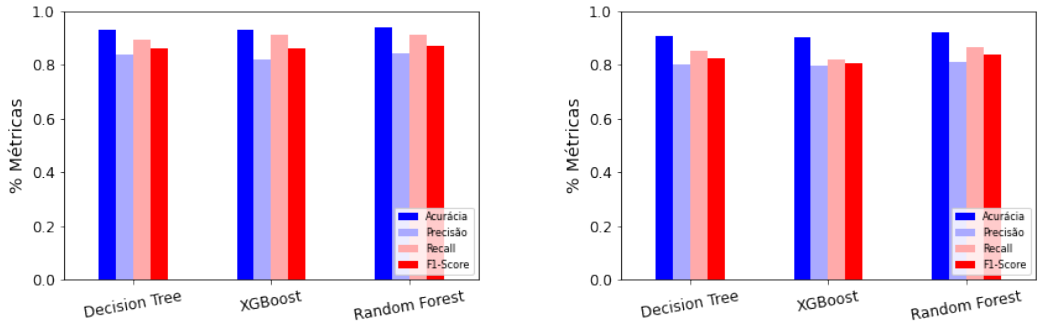
Figura 4. Fronteira de Pareto com a seleção das características com maior correlação e maior informação mútua referentes ao Cenário 3. As características que se encontram nesta fronteira não são dominadas por nenhuma outra. a) Pares de características escolhidos para o dia 26/02/2017. b) Pares de características escolhidos características escolhidas para o dia 03/02/2017.

Para o Cenário 4 foram utilizados todos os pares de características obtidos pela Fronteira de Pareto do Cenário 3, referentes ao dia 26/02/2017. Entretanto, os pares de características foram interpretados como arestas em um grafo não conexo. Assim, a cada componente conexa do grafo de características, foi aplicado o conceito de centralidade para diminuir o número de características selecionadas. A centralidade avalia o grau de importância que cada vértice possui em um grafo. Como cada vértice representa uma característica e, para esse cenário, as características possuem alta informação mútua e alta correlação de Pearson, a centralidade avalia quais dessas características possuem mais informação compartilhada com as demais características. Dentre as 11 características selecionadas pela Fronteira de Pareto, após a aplicação da seleção pelas características mais centrais nas componentes conexas, foram obtidas 4 características (*total-bhlen, sflow-fbytes, total-fpackets, max-fpktl*) que representam a quantidade total de bytes utilizados por cabeçalhos na direção oposta, o número médio de bytes transmitidos em um sub-fluxo na direção para frente, número de pacotes transmitidos para frente e o tamanho do maior pacote em bytes transmitido para frente.

A Figura 5 apresenta os valores das métricas Acurácia, Precisão, Revocação e *F1-Score*, considerando os três classificadores distintos: Árvore de Decisão (`Decision Tree`), `XGBoost` e Floresta Aleatória (`Random Forest`). O Cenário 2, utilizando as características que minimizam Informação Mútua e Correlação de Pearson, apresenta alta acurácia, porém indica perda de revogação em relação aos outros cenários. Nesse cenário, foram utilizadas 7 características. Tal fenômeno indica que a seleção de características pouco correlacionadas e com baixa informação mútua captura informação relevante da classe mais abundante no conjunto de dados, porém falha ao avaliar a classificação da classe menos abundante. A abordagem que utilizou o menor número de características



(a) Cenário 1 (b) Cenário 2



(c) Cenário 3 (d) Cenário 4

Figura 5. Comparativo entre métodos de seleção de características. a) Classificação utilizando todas as características. b) Minimização conjunta. c) Maximização conjunta. d) Seleção das características centrais.

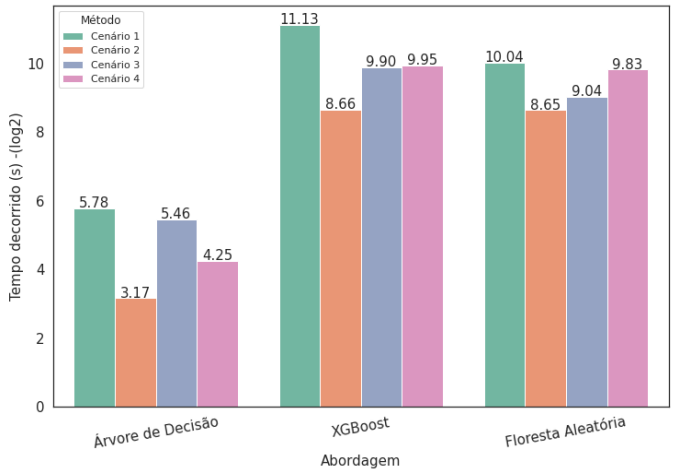


Figura 6. Comparativo entre o tempo de treinamento e validação para os métodos propostos. O gráfico está em escala logarítmica.

para o treinamento foi o Cenário 4. No total, foram utilizadas apenas 4 características. Um pequeno número de características utilizadas para treinamento e validação, com resultados similares aos outros métodos, indica que as características utilizadas possuem um alto grau de informação sobre o conjunto de dados. Outro aspecto observado refere-

se ao tempo de treinamento e de validação do conjunto de dados. A Figura 6 apresenta o tempo em segundos que cada modelo levou para treinar e validar a classificação. A escala logarítmica foi empregada em função da alta diferença de tempo entre os demais modelos e modelo Árvore de Decisão. O modelo da Árvore de Decisão apresenta o melhor desempenho em todos os cenários, mantendo Precisão e Acurácia próxima aos demais. Por fim, observa-se que os Cenários 2 e 4 são os que mais reduzem a redundância de características selecionadas e, portanto, são os que levam a modelos mais gerais. Assim, embora haja uma pequena degradação de desempenho ao se comparar com o modelo composto por todas as características, os Cenários 2 e 4 utilizam menos recursos computacionais e reduzem a probabilidade de sobreajuste.

6. Conclusão

A seleção de características para treinamento de classificadores na identificação de ameaças em redes é um desafio. Com o aumento no compartilhamento de recursos computacionais, torna-se necessária uma análise mais detalhada em fluxos de dados para a identificação de comportamento anormal na rede. A necessidade de modelos de predição e classificação que possam ser executados em dispositivos com baixo poder computacional é fundamental. No entanto, treinar modelos de aprendizado de máquinas possui o desafio de encontrar um subconjunto ótimo de características em um grande conjunto de dados. Este artigo apresentou uma proposta para seleção de características utilizando correlação de Pearson, Informação Mútua e Fronteira de Pareto. Para avaliar a proposta, foi utilizado um conjunto de dados de um grande provedor de telecomunicações, coletado na cidade do Rio de Janeiro. Foram calculadas as correlações de Pearson e Informação Mútua, para pares de características do conjunto de dados. Em seguida, após o processamento, foram avaliados quatro cenários, no qual, as características selecionadas pelo método de minimização da informação mútua e da correlação obtiveram o melhor desempenho quanto à quantidade de características utilizadas, mantendo desempenho similar em acurácia, precisão e revogação ao conjunto completo de todas as características. O algoritmo de Árvore Aleatória apresentou o melhor desempenho, sobretudo no tempo de treinamento e classificação. Como trabalhos futuros, pretende-se avaliar a proposta em novos conjuntos de dados.

Agradecimentos

Este trabalho foi realizado com recursos do CNPq, FAPERJ, RNP, CAPES, CGI/FAPESP (2018/23062-5) e Prefeitura de Niterói/FEC/UFF (Edital PDPA 2020).

Referências

Abdollahzadeh, B. e Gharehchopogh, F. S. (2022). A multi-objective optimization algorithm for feature selection problems. *Engineering with Computers*, 38(3):1845–1863.

Andreoni Lopez, M., Mattos, D. M. F., Duarte, O. C. M. B. e Pujolle, G. (2019). A fast unsupervised preprocessing method for network monitoring. *Annals of Telecommunications*, 74(3):139–155.

Andreoni Lopez, M., Sanz, I. J. e Lobato, A. G. P. (2018). Aprendizado de máquina em plataformas de processamento distribuído de fluxo: Análise e detecção de ameaças em tempo real. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) - Minicursos*.

Arifeen, M., Petrovski, A. e Petrovski, S. (2021). Automated microsegmentation for lateral movement prevention in industrial internet of things (iiot). Em *2021 14th International Conference on Security of Information and Networks (SIN)*.

Di Mauro, M., Galatro, G., Fortino, G. e Liotta, A. (2021). Supervised feature selection techniques in network intrusion detection: A critical review. *Engineering Applications of Artificial Intelligence*, 101:104216.

Farrugia, S., Ellul, J. e Azzopardi, G. (2020). Detection of illicit accounts over the ethereum blockchain. *Expert Systems with Applications*, 150:113318.

Garg, S., Kaur, K., Kumar, N., Kaddoum, G., Zomaya, A. Y. e Ranjan, R. (2019). A hybrid deep learning-based model for anomaly detection in cloud datacenter networks. *IEEE Transactions on Network and Service Management*, 16(3):924–935.

Kasongo, S. M. e Sun, Y. (2019). A deep learning method with filter based feature engineering for wireless intrusion detection system. *IEEE Access*, 7:38597–38607.

Kim, T.-Y. e Cho, S.-B. (2018). Web traffic anomaly detection using c-lstm neural networks. *Expert Systems with Applications*, 106:66–76.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. e Liu, H. (2017). Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6).

Lopez, M. A., Silva, R. S., Alvarenga, I. D., Rebello, G. A. F., Sanz, I. J., Lobato, A. G. P., Mattos, D. M. F., Duarte, O. C. M. B. e Pujolle, G. (2017). Collecting and characterizing a real broadband access network traffic dataset. Em *2017 1st Cyber Security in Networking Conference (CSNet)*, p. 1–8.

Ma, Q., Sun, C., Cui, B. e Jin, X. (2021). A novel model for anomaly detection in network traffic based on kernel support vector machine. *Computers & Security*, 104:102215.

Matin, I. M. M. e Rahardjo, B. (2019). Malware detection using honeypot and machine learning. Em *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, volume 7, p. 1–4.

Medeiros, D., Cunha Neto, H., Andreoni, M., Magalhães, L., Silva, E., Borges, A., Fernandes, N. e Menezes, D. (2019). *Análise de Dados em Redes Sem Fio de Grande Porte: Processamento em Fluxo em Tempo Real, Tendências e Desafios*, p. 142–195.

Silva, J. V. V., de Oliveira, N. R., Medeiros, D. S., Lopez, M. A. e Mattos, D. M. (2022). A statistical analysis of intrinsic bias of network security datasets for training machine learning mechanisms. *Annals of Telecommunications*, p. 1–17.

Thakkar, A. e Lohiya, R. (2021). Attack classification using feature selection techniques: a comparative study. *Journal of Ambient Intelligence and Humanized Computing*, 12(1):1249–1266.

Viduto, V., Maple, C., Huang, W. e López-Peréz, D. (2012). A novel risk assessment and optimisation model for a multi-objective network security countermeasure selection problem. *Decision Support Systems*, 53(3):599–610.

Wang, W., Liang, C., Chen, Q., Tang, L., Yanikomeroglu, H. e Liu, T. (2022). Distributed online anomaly detection for virtualized network slicing environment. *IEEE Transactions on Vehicular Technology*.