Técnicas de reconhecimento de padrões para identificação de ataque de DNS

Henrique Carlos Fonte Boa Carvalho [1]
Eduardo Pelli [1]

Resumo: A maioria dos dispositivos conectados à Internet usufrui dos serviços de DNS (Domain Name System) para resolução de nomes de domínios. A partir da afirmação de que a maioria das redes não restringe o tráfego de pacotes destinados aos serviços de DNS, técnicas de ataques podem ser aplicadas, fazendo com que uma simples requisição de resolução de nome aparentemente normal, possa ocasionar um ataque causando diversos transtornos às vítimas. Os ataques do tipo Spoofing consistem em enganar o dispositivo do usuário, fazendo com que o computador identifique o dispositivo do usuário malicioso, de maneira confiável. Por se tratar de um tipo de ataque de alta periculosidade, devido à inexistência de pesquisas e técnicas eficientes na sua identificação, procura-se encontrar soluções viáveis para detecção deste tipo de ataque. Este trabalho teve como objetivo a aplicação de técnicas de Reconhecimento de Padrões através da aplicação das técnicas de Seleção de Características e Classificação de características para detecção de DNS Spoofing em redes locais de computadores. Foram obtidos resultados com acurácia média de 98,33% $\pm$ 0,64% na detecção na classe de falha da rede, ou seja, quando essas estavam sob ataque de DNS Spoofing.

***Palavras-chave*:** Reconhecimento de Padrões. Segurança da Informação. Spoofing. SVM.

Abstract: *Most devices connected to the Internet use the DNS service (Domain Name System) to resolve domain names. From the statement that most networks do not restrict packet traffic destined for DNS services, technical attacks can be applied, making a simple request apparently normal name resolution, can lead to an attack causing several disorders to victims. The type of Spoofing attacks consist in fooling the user's device, causing the computer to identify the malicious user device reliably. Because it is a type of highly dangerous attack, due to the lack of research and efficient techniques in identifying, looking to find viable solutions to detect this type of attack. This study aimed to apply pattern recognition techniques by applying the techniques of selection characteristics and classification features for DNS Spoofing detection in local computer networks. Results were obtained with an average accuracy of 98,33% $\pm$ 0,64% in detecting the failure of the network element, or when these were under DNS spoofing attack.*

***Keywords*:** *Information Security. Pattern Recognition. Spoofing. SVM*

1 Introdução

O número de usuários que trocam informações na Internet está cada vez maior. Em quinze anos, passou de 400 milhões para 3,2 bilhões [1]. Cerca de 42,5 milhões de brasileiros utilizam celulares para acessar a Internet, sendo que o número de usuários que o fazem no país é de 85,9 milhões [2]. Com esse crescente aumento de usuários utilizando computadores e dispositivos móveis para acessar redes locais e a Internet, uma maior quantidade de dados e informações é armazenada e transmitida entre usuários.

Com o aumento de usuários, tráfego de dados e da disseminação da informação nas redes de computadores, vem crescendo o número de usuários maliciosos que possuem o objetivo de roubar dados e informações. Eles buscam muitas vezes benefícios próprios ou apenas semear o caos e são conhecidos como *crackers* [3, 4].

[1]Departamento de Computação, UFVJM, Campus JK - Rodovia MGT 367 - Km 583, nº 5000 - Diamantina (MG) - Brasil
{henrique.fonteboa,pellie@ufvjm.edu.br}

http://dx.doi.org/10.5335/rbca.v9i2.6279

Uma técnica que foi criada para obtenção de dados privilegiados, ou seja, dados que apenas outros usuários possuem permissão para acessar, foi o *Spoofing*. Ele consiste basicamente, em fornecer uma credencial falsa ao destinatário para, através de um conjunto de outros métodos, garantir que consiga obter os dados desejados. Dessa maneira, aparentemente, toda conexão realizada foi enviada e recebida corretamente sem que nenhum usuário malicioso tenha acesso. Essa técnica é de alta periculosidade, não apresentando métricas eficazes e defesas definitivas para sua identificação.

As técnicas de Reconhecimento de Padrões podem ser utilizadas com o intuito de identificar padrões, ou seja, estudam um grupo de dados e buscam maneiras de agrupá-las e classificá-las através de padrões identificados [5].

O objetivo deste trabalho é identificar as características relevantes em um ataque de *DNS Spoofing* e, por meio de suas características, identificar novos ataques. O presente trabalho foi dividido da seguinte maneira: a Seção 2 define os principais conceitos utilizados no trabalho; a Seção 3 apresenta os trabalhos relacionados; a Seção 4 detalha a metodologia utilizada; a Seção 5 apresenta os resultados obtidos; a Seção 6 apresenta as considerações finais.

2 Conceitos Utilizados

2.1 Spoofing

O *Spoofing* é uma técnica que consiste em fornecer uma identidade falsa, que por sua vez é considerada confiável aos usuários alvos que estão conectados à rede. Essa técnica consegue, por meio da identidade falsa, modificar e enviar pacotes falsos ou ilegais aos alvos. O *Spoofing* possui algumas variações no enfoque do ataque, podendo ser usado basicamente: o *IP Spoofing*, o *DNS Spoofing* e também o *ARP Spoofing*.

O *IP Spoofing* consiste em enganar o dispositivo do usuário alvo, fazendo com que o computador alvo identifique o dispositivo do usuário malicioso de maneira confiável. Desta forma, o dispositivo alvo acaba por confiar e receber qualquer tipo de pacote de informações provenientes do atacante. Com sua utilização, é possível que o atacante invada o computador do usuário alvo, receba os pacotes provenientes da comunicação entre os dispositivos e cometa um *hijacking*, que é um sequestro da seção do usuário.

O IP spoofing é uma das maneiras mais comuns de se camuflar na internet. Através dessa técnica, o atacante ganha um acesso não autorizado a um computador ou uma rede por fazer parecer que a mensagem foi enviada através de uma máquina confiável, pois o endereço de IP foi falsificado [6].

O *ARP Spoofing* é um ataque cujo objetivo é alterar a resposta *ARP* enviada a uma requisição original através de uma resposta falsa. Enviando uma resposta falsa, o roteador pode expedir dados destinados ao computador da vítima para o computador do usuário malicioso, e este redireciona os dados para a vítima. Se for bem sucedido, o computador da vítima não tem ideia que o redirecionamento das informações está ocorrendo.

O *DNS Spoofing* consiste em enganar o dispositivo do usuário alvo, conseguindo redirecionar as requisições realizadas ao servidor de DNS dos aplicativos Web que o dispositivo alvo deseja acessar, propiciando um redirecionamento para uma página definida pelo usuário malicioso. O *DNS Spoofing* consiste de basicamente quatro etapas:
1ª O usuário alvo realiza uma requisição ao servidor de DNS;
2ª O invasor retorna um IP falso para a requisição do usuário alvo antes de chegar a resposta do servidor de DNS com o IP verdadeiro;
3ª A resposta do servidor de DNS com o IP verdadeiro é recebida e descartada pelo alvo, pois já foi recebida uma resposta anteriormente;
4ª O usuário acessa o endereço fornecido pelo invasor;

O ataque de *DNS Spoofing* exemplificado anteriormente pode ser observado na Figura 1.

Há três mecanismos utilizados atualmente para bloqueio de pacotes falsos provenientes de *IP Spoofing*, são eles: filtragem de pacotes que chegam e saem da rede origem (Ingresso / Egresso); rastreamento inverso (*Trace back*); e tentativas de descarte de pacotes no destino. [7] propuseram um método de prevenção do *Ip Spoofing*

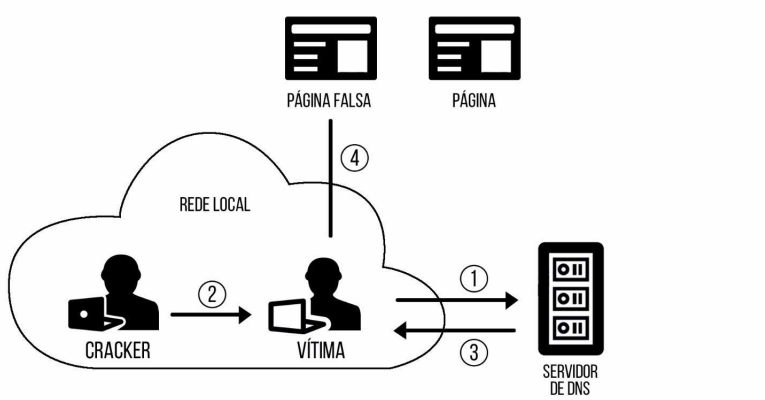
(Spoofing Prevention Method - SPM), com precisão de 99%. Tal método se baseia na análise de autenticidade do endereço da fonte de pacotes por roteadores próximos aos dispositivos destinos destes pacotes. A avaliação é dada de uma chave única associada a um registro temporal para cada conexão de rede com origem e destino definidos.

Essas soluções não se aplicam ao *DNS Spoofing*, pois atuam nas redes locais, não sendo possível a filtragem dos pacotes por um roteador ou *Firewall*. Portanto, esse tipo de ataque é considerado de alta periculosidade, pois, não existem técnicas eficientes para sua identificação, sendo necessário o estudo de todas as variáveis relacionadas ao ataque, a fim de se tentar criar estratégias de defesa.

Ataques de *DNS Spoofing* são realizados em dispositivos localizados em locais públicos, em busca de obtenção de informação sobre cartões de crédito de bancos. Para prevenir um ataque deste tipo, deve-se atentar ao contexto e às decisões de segurança [8].

De acordo com [9], em 2005 foram contabilizados pelo menos quatro mil ataques de *Spoofing* a cada semana na *Internet*, devido à facilidade com que o ataque é gerado. Atualmente, com um nível de acesso muito maior, pode-se presumir que o nível de ataque é bastante superior.

Figura 1: Esquema representativo de um ataque de *DNS Spoofing*.



Fonte: Elaborado pelo autor.

2.2 Reconhecimento de Padrões

De acordo com [10], reconhecimento de padrões é a área que estuda como as máquinas podem observar dados, aprender e distinguir os dados fornecidos para então classificar e categorizar os dados em classes através de características comuns.

Existem duas formas de reconhecer e classificar um padrão: o reconhecimento supervisionado, quando os dados fornecidos já foram identificados com uma classe fornecida; e o reconhecimento não supervisionado, quando não há informações fornecidas previamente por um agente externo. Assim, o sistema não apresenta informações conhecidas como dados de entrada [11].

Na primeira etapa, define-se como e quais os dados serão obtidos para, posterior análise. Nesta etapa ocorre a extração de características, além de ocorrer o pré-processamento dos dados, que consiste em aplicar determinadas técnicas visando a obter ajustes na base de dados. Desta maneira, é possível aumentar ou diminuir algumas características.

Na segunda etapa são utilizadas técnicas de seleção de características para identificar quais características da base de dados são mais relevantes. Os dados são, em seguida classificados e agrupados em classes. Para tarefa de seleção, pode-se citar os métodos *F-Score* e Coeficiente de Correlação de Pearson.

Na terceira etapa, é realizada a classificação, a qual consiste em utilizar as características que foram obtidas nas etapas anteriores para conseguir diferenciar novos dados fornecidos em classes.

2.3 Seleção de Características

Segundo [5], o método *F-Score* (*Fisher Score*) para seleção de característica disponibiliza uma medida dada pela distância entre as médias das distribuições de duas classes (C_1 e C_2) em relação às suas variâncias. Quanto maior é o valor da classe calculado pelo *F-Score*, mais discriminativa e relevante é a característica para o experimento. A métrica é definida pela Equação 1 [12]:

$$F(g) = \frac{\sum_{k=1}^{n} n_k(\mu_k^j - \mu^j)^2}{(\sigma^j)^2} \tag{1}$$

Onde: $(\sigma^j)^2 = \sum_{i=1}^{n} n_k(\sigma_k^j)^2$; *F(g)* é a função que calcula o valor *F-score* para a característica *g*;

O Coeficiente de Correlação de *Pearson* é outro método para ranquear características relevantes. Ele mede o grau de relação da distribuição de duas classes. O cálculo desse coeficiente é definido pela Equação 2 [5]:

$$\rho_j = \frac{\frac{1}{n-1}\sum_{i=0}^{n}(x_{ij} - \bar{x}) \times (y_{ij} - \bar{y})}{\sigma_{xj} \times \sigma_y} \tag{2}$$

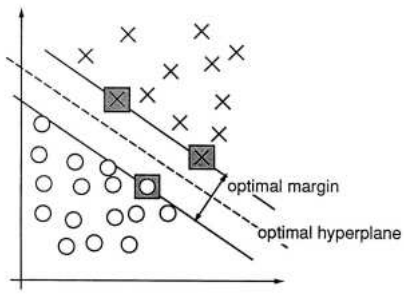
Onde: -1 $\leq \rho_j \leq$ 1; Se $\rho_j = 1$, existe total relação positiva entre as distribuições; Se $\rho_j = -1$, existe total relação negativa entre as distribuições; Se $\rho_j = 0$, as distribuições não possuem relação.

2.4 SVM

O classificador SVM foi desenvolvido baseado nas ideias originadas na teoria de aprendizagem estatística [13]. De acordo com [11], a teoria de aprendizagem estatística estabelece uma série de métricas que devem ser seguidas, com o intuito de obter um classificador com boa generalização, definida como a sua capacidade de identificar corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu.

Segundo [14], a SVM pode ser descrita da seguinte maneira: determinada base da dados com duas classes e um conjunto de dados pertencentes a uma dessas classes, a SVM encontra o hiperplano que separa os dados de ambas as classes (Figura 2). O hiperplano deve cometer poucos erros marginais, minimizando assim o erro sobre os dados de teste e de treinamento, respectivamente. O hiperplano, então, é denominado ótimo. O hiperplano é determinado por uma amostra da base de dados, nomeado vetores de suporte. O conceito por traz do SVM é a maximização da margem, ou seja, maximizar a distância dos dados de treinamento [11], [15].

Figura 2: Distribuição do hiperplano.



Fonte: Adaptado de Cortes [16].

A decisão pelo *SVM* é dado pela (Equação 3) [16].

$$W(\alpha) = \sum_{i}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{3}$$

$$Onde = \sum_{i=1}^{N} y_i \alpha_i; \forall_{i=1}^{N} : 0 \leq \alpha_i \leq C \quad (4)$$

Onde: $W(\alpha)$ é o vetor de pesos; $K(x_i, x_j)$ é o Kernel utilizado; α é o multiplicador de Lagrange; C é especificado pelo usuário; X é o vetor de características; Y é o rótulo da classe;

O parâmetro C é a relação entre a complexidade do algoritmo e o número de amostras de treinamento incorreto [17]. De acordo com [18], pode ser visto como um parâmetro de penalização.

As funções *kernel* são o que diferenciam um modelo de SVM de outro, ou seja, a função realiza o mapeamento dos dados [19]. Segundo [20], o *kernel* realiza implicitamente o mapeamento nos dados de características e depois usa um algoritmo para classificar o espaço.

O *SVM* possui quatro funções *kernel*:

- Linear = $(x' * y)$
- Polinomial = $(\gamma * x' * v + 1)^d$
- Radial Basis = $\exp(-\gamma * |u - v|^2)$
- Sigmoidal = $\tanh(\gamma * u' * v)$

3 Trabalhos Relacionados

Até o presente momento, não foram identificados na literatura trabalhos que utilizem técnicas de reconhecimento de padrões como *SVM* para identificação de características relevantes para um ataque de *DNS Spoofing*.

Trabalhos relacionados à aplicação do *SVM*, vêm sendo amplamente utilizados em diversas áreas do conhecimento. Trabalhos de destaque podem ser encontrados na categorização de textos, segundo [21]; na análise de imagens, segundo [22] e [23]; e em bioinformática, de acordo com [24] e [25].

Em [26], através da utilização do SVM, obteve uma acurácia taxa de acerto para a predição de novos dados de 89,13%. No estudo, utilizou-se o classificador para identificar motoqueiros que estavam com ou sem capacete através da análise de imagens, e foi obtida uma acurácia de 81,37%. Em [27], na utilização do SVM para classificação no mapeamento do uso e cobertura da terra na Amazônia através de imagens de satélite e [28] obteve uma acurácia de aproximadamente 75%, sendo considerado no estudo um desempenho muito bom. Seu estudo foi para a classificação de imagens de satélite na região da bacia de drenagem do reservatório de Itumbiara e apresentou-se superior a outros métodos de classificação [29], [30] e [31].

4 Metodologia

Após revisão bibliográfica, foi realizada a extração das características (Seção 4.1). Posteriormente foi realizada a coleta de dados (Seção 4.2), e, em seguida, o pré-processamento dos dados (Seção 4.3), posteriormente o pré-processamento foi realizado a seleção das características (Seção 4.5). Com as características mais relevantes foram definidos quais experimentos seriam realizados, foi realizado a classificação das características (Seção 4.5) e avaliado o resultado do classificador (Seção 4.6).

4.1 Extração de características

Após a etapa de leitura bibliográfica ocorreu a etapa de Extração de Características, na qual realizaram-se testes e análises de verificação características que poderiam estar relacionadas a um ataque de *DNS Spoofing*. Foram identificadas as seguintes características: Saltos, Tempo de Resposta, Erros.

- Saltos: basicamente é a quantidade de *hops* que o pacote demora para chegar ao destino, ou seja, a quantidade de dispositivos da rede que um pacote passa até chegar ao servidor de destino. Também conhecido na literatura por TTL (*Time To Live*), foram identificados através do comando *traceroute*, que envia pequenos pacotes, por cada nó da rede até que o pacote chegue ao destino. Posteriormente ele retorna para a máquina solicitante. Para obtenção de uma grande quantidade de dados é despendida uma enorme quantidade de horas, pois precisa aguardar uma resposta de um nó para enviar o novo pacote para o próximo nó da rede.
- Tempo de Resposta: é o tempo de resposta entre dois pontos de uma rede. Também conhecido na literatura como RTT *(Round Trip Time)*, foi obtido através do comando *ping*, que envia um pacote a um dispositivo conectado à rede local ou global e aguarda sua resposta. O tempo de resposta é calculado em milissegundos. Quando determinado pacote que foi enviado e o pacote de retorno apresenta um grande atraso, é descartado sem apresentar um tempo de resposta. Para a obtenção do tempo de resposta, não é necessário um elevado poder de processamento e nem despendido muito tempo.
- Erros: quando um pacote passa do tempo máximo permitido pelo comando *ping*, é considerado como um "erro"e é descartado pelo comando, não apresentando tempo de resposta. Desta maneira, quando o pacote era descartado, o valor do tempo de resposta foi considerado como 2000 ms; erros = 1.

Para a pesquisa, foram selecionadas as características Tempo de Resposta e Erros, pois as mesmas estão diretamente relacionadas e ambas as características podem ser obtidas com um baixo custo computacional e temporal em comparação com o número de Saltos.

4.2 Coleta de Dados

Após identificadas as características envolvidas a um ataque de *DNS Spoofing*, foi desenvolvido um *script* para a obtenção dos dados. O script foi aplicado em dois momentos: durante uma ocorrência de ataque, em que foram coletados 2500 dados de cada característica; e também foi aplicado durante nenhuma ocorrência de ataque na rede, sendo obtidos também 2500 dados, quando foi armazenando os dados referentes às características: Tempo de Resposta e Erros. Por último, foi acrescentada uma característica denominada Classe, que é responsável por realizar o rótulo correto de cada amostra. O rótulo "1" foi definido para os dados que estavam sofrendo um ataque de *DNS spoofing* e "2" para os que não estavam sofrendo nenhum tipo de ataque.

4.3 Pré-processamento dos Dados

De posse dos 5 mil dados obtidos, foi realizado o pré-processamento dos dados. Foram obtidos dados mais homogêneos sem grande dispersão por possíveis erros.

Para a realização desta etapa foi desenvolvido um *script* em Linguagem R. Cada característica foi transformada em sua média e seu desvio padrão. Para obtenção desses dados, era calculada a média ou desvio padrão a cada dez dados. Após o cálculo realizado nas características, foi obtida uma nova base de dados, contendo quatro características e 500 repetições de cada.

As novas características foram: Média do Tempo de Resposta, Desvio padrão do Tempo de Resposta, Média de Erros e Desvio padrão de Erros.

4.4 Seleção de Características

Após o ajuste realizado na base de dados que continha os dados originais, foram aplicadas as técnicas de seleção de características como *F-score* e o *Coeficiente de Correlação de Pearson*, com o objetivo de identificar as características mais relevantes para o estudo. Durante a aplicação das técnicas, foi utilizada a totalidade dos dados, ou seja, todas as quatro características e suas 500 repetições.

Com a aplicação das técnicas, foi obtida uma tabela de ranqueamento dos dados ordenados pelo valor de cada uma, onde são apresentadas as características relevantes para o estudo.

Após a aplicação da seleção, foi realizada uma alteração nas Classes dos dados, ou seja, foi alterado o rótulo dos dados, para o resultado ser mais claro e mais conciso, desta forma, os dados anteriores que possuíam o rótulo

de “1” passaram para “S”, para simbolizar que estava sendo realizado um ataque de *Spoofing*. Já os que possuíam o rótulo de “2” passaram para “N”, onde não ocorria nenhum tipo de ataque.

4.5 Classificação de Características

De posse dos dados e do resultado de quais são as características mais relevantes, foi definido quais seriam os experimentos realizados, para posteriormente aplicar o classificador *SVM* com as características escolhidas. O aprendizado escolhido para esse classificador foi o supervisionado e, neste caso, era necessário fornecer ao *SVM* os dados para o seu “treinamento” e “teste”.

A base de dados fornecida para o “treinamento” continha as características e apresentava o rótulo de cada dado, ou seja, apresenta os dados e mostra se o dado foi proveniente de um ataque de *Spoofing* ou não. A base de “treinamento” possui a função de fornecer conhecimento para o classificador. Para o “teste”, foi utilizado o restante dos dados, sem a utilização dos rótulos. Eles não estão contidos na base de “treinamento” e, desta maneira, a base de “teste” era uma base de dados “nova”, sendo que seus dados não foram fornecidos previamente ao classificador. A base de “teste” é utilizada para verificar qual o percentual de acerto do classificador para novos dados; assim sendo, é utilizada para verificar se o classificador consegue prever corretamente a qual classe os dados desconhecidos pertencem.

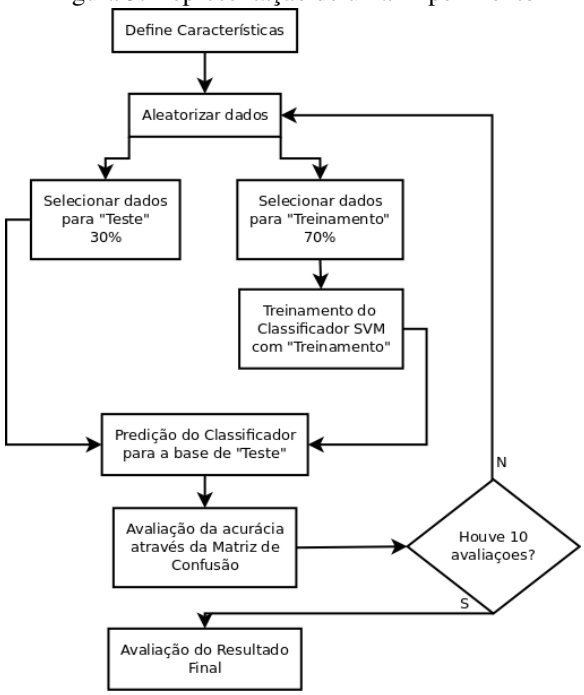
Com o intuito de obter dados imparciais e não tendenciosos durante a classificação e a avaliação dos resultados da taxa de acerto do classificador SVM, foi aplicado o classificador por dez vezes para obtenção da média final dos resultados para cada experimento. Para cada repetição do classificador sobre um experimento, foram utilizados dados diferentes para “treinamento” e para “teste”, a partir dos quais eram fornecidos dados randômicos da base de dados. A base de dados de “treinamento” possuía 70% e a de “teste” possuía 30% dos dados. Toda aplicação da SVM foi realizada através da utilização do *Kernel* linear.

Foram realizados três experimentos diferentes com a aplicação do classificador SVM para identificar com quais características ele apresentaria uma maior acurácia, ou seja, uma maior taxa de acerto para novas entradas, da seguinte forma.

- Experimento 1 - SVM para as 4 características
- Experimento 2 - SVM para Média do Tempo de Resposta e Desvio Padrão do Tempo de Resposta
- Experimento 3 - SVM para as Média de Erros e Desvio Padrão de Erros

A Figura 3 apresenta o que ocorreu em cada experimento. Ela exibe desde a escolha das características utilizadas para a aplicação do classificador até a etapa de avaliação dos resultados. Para chegar à avaliação dos resultados, foram realizadas dez repetições do classificador contendo as mesmas características e avaliando o resultado da matriz de confusão, para posteriormente calcular a média dos valores.

Figura 3: Representação de uma Experimento



Fonte: Elaborado pelo autor.

4.6 Métrica de avaliação dos resultados

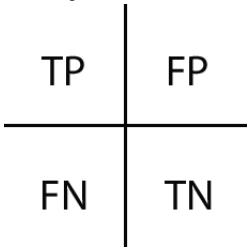
Foi avaliada a precisão do classificador após dez repetições sobre cada base de dados composta pela média e o desvio padrão das características originais. Para analisar a precisão do nosso classificador de forma não tendenciosa, a cada repetição eram utilizados novos dados para "treinamento" e "teste". Por conseguinte, todas as avaliações do classificador foram obtidos através de dados diferentes, acarretando resultados diferentes.

Após o "treinamento" dos dados, foi realizada a predição pelo classificador onde ele tentava identificar a qual classe os dados de "teste" pertencem, e, pelos resultados do classificador *SVM*, foi gerando a Matriz de Confusão.

Segundo [32], a matriz de confusão oferece métricas efetivas de identificar se as predições realizadas pelo classificador foram corretas ou incorretas. A matriz de confusão pode ser representada segundo a Figura 4. Através da análise da matriz podem-se obter as seguintes informações:

- Acurácia: Taxa (%) das amostras que foram classificados corretamente.
- Precisão: Taxa (%) com que o classificador oferece resultados similares após
- Sensibilidade: Taxa (%) de classificação positiva das amostras realmente positivas.
- Especificidade: Taxa (%) de classificação negativa das amostras realmente

Figura 4: Representação de uma Matriz de Confusão



Fonte: Adaptado de Senna [32].

Onde:

- TP: True Positive ou Verdadeiro Positivo
- FP: False Positive ou Falso Positivo.
- FN: False Negative ou Falso Negativo.
- TN: True Negative ou Verdadeiro Negativo.

Foram obtidos acurácia, precisão, sensibilidade e especifidade, porém, neste trabalho, a característica relevante foi a acurácia. Os resultados analisados foram obtidos através das seguintes Equações:

$$Acurcia = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$Preciso = \frac{TP}{TP + FP} \tag{6}$$

$$Sensibilidade = \frac{TP}{|P|} = \frac{TP}{TP + FN} \tag{7}$$

$$Especifidade = \frac{TN}{|N|} = \frac{TN}{TN + FP} \tag{8}$$

5 Resultado e Discussão

Os resultados foram divididos em quatro subseções, nas quais serão apresentados resultados e posteriormente será realizada uma consideração sobre os resultados.

5.1 Seleção de Características

Os resultados obtidos através da métrica de seleção de características *F-Score* é exibido (Tabela 1).

Tabela 1: Características ranqueadas através do F-Score

Característica	Nome	F-score
1	mediatempoderesposta	0.1539
2	stdtempoderesposta	0.0423
4	stderros	0.0094
3	mediaerros	0.0074

Fonte: Elaborado pelo autor.

Os resultados obtidos através da métrica de seleção de características Coeficiente de Correlação de Pearson (Tabela 2).

Tabela 2: Características ranqueadas através do Coeficiente de correlação de Pearson

Característica	Nome	Pearson
1	mediatempoderesposta	0.3659
2	stdtempoderesposta	0.2020
4	stderros	0.0970
3	mediaerros	0.0861

Fonte: Elaborado pelo autor.

Observa-se que os resultados apresentados pelas duas métricas de seleção de características F-Score e Coeficiente de Correlação de Pearson apresentaram a mesma ordem de relevância para as características.

A característica que apresentou maior relevância para o estudo foi Média Tempo de Resposta, através do F-Score, foi obtido 0,1539. Através do Coeficiente de Correlação de Pearson, a característica obteve o valor de 0,3659.

5.2 Experimento 1 - SVM para as 4 características

O resultado da média e desvio padrão das dez repetições foram obtidos pelas avaliações da aplicação do classificador *SVM*, utilizando-se a base de dados com as quatro características (Tabela 3).

Tabela 3: Resultado do classificador para as 4 características

	Acurácia	Precisão	Sensibilidade	Especifidade
Média	98,33%	96,77%	100%	96,58%
Desvio padrão	0,64%	1,28%	0,00%	1,39%

Fonte: Elaborado pelo autor.

Observa-se que quando o classificador SVM foi treinado e testado com as quatro características utilizadas no estudo, apresentou uma acurácia média de 0,9833, ou seja, conseguiu identificar corretamente 98,33% dos dados corretamente, apresentando uma elevada taxa de acerto.

5.3 Experimento 2 - SVM para Média do Tempo de Resposta e Desvio Padrão do Tempo de Resposta

O resultado da média e desvio padrão das dez repetições foram obtidos através das avaliações da aplicação do classificador *SVM*, utilizando-se a base de dados com as características Média do Tempo de Resposta e Desvio Padrão do Tempo de Resposta (Tabela 4).

Tabela 4: Resultado do classificador para Média Saltos e Média Tempo de Resposta

	Acurácia	Precisão	Sensibilidade	Especifidade
Média	97,67%	96,03%	99,46%	95,85%
Desvio padrão	1,00%	1,73%	0,69%	1,80%

Fonte: Elaborado pelo autor.

Observa-se que foi obtida uma acurácia média de 0,9767, ou seja, conseguiu-se identificar corretamente 97,67% de novos ataques de *DNS Spoofing*.

5.4 Experimento 3 - SVM para as características Média de Erros e Desvio Padrão de Erros

O resultado da média e desvio padrão das dez repetições foram obtidos através das avaliações da aplicação do classificador *SVM*, utilizando-se a base de dados com as as características Média de Erros e Desvio Padrão de Erros (Tabela 5).

Foi obtido uma acurácia de 0,5300, ou seja, conseguiu identificar corretamente 53,00% de novos ataques corretamente.

Tabela 5: Resultado do classificador para as características Média de Erros e Desvio Padrão de Erros

	Acurácia	Precisão	Sensibilidade	Especifidade
Média	53,00%	57,67%	42,50%	66,43%
Desvio padrão	6,84%	11,17%	30,74%	35,15%

Fonte: Elaborado pelo autor.

6 Conclusão

Através deste trabalho, observa-se que a aplicação de reconhecimento de padrões com a aplicação de técnicas de seleção de características como F-Score, Coeficiente de correlação de Pearson e a técnica de classificação de características como *SVM*, é possível predizer com 98,33% de acurácia novos ataques de *DNS Spoofing*.

A característica Média do Tempo de Resposta obteve o melhor resultado através das técnicas de seleção de características, sendo considerada a mais relevante para o estudo. Através da utilização das características Média do Tempo de Resposta e Desvio Padrão do Tempo de Resposta, foram obtidos os melhores resultados, sendo possível predizer com alta taxa de acerto os casos de novos ataques.

A característica Erros, sozinha, possui uma baixa taxa de acerto para novos ataques de *DNS Spoofing*, predizendo corretamente apenas 53% de novos ataques, mas em conjunto com o Tempo de Resposta, apresenta uma elevada acurácia de 98,33%, apontando o melhor experimento realizado.

Referências

[1] ONUBR. *Em 15 anos, número de usuários de internet passou de 400 milhões para 3,2 bilhões, revela ONU*. 2015. Disponível em: <https://nacoesunidas.org/em-15-anos-numero-de-usuarios-de-internet-passou-de-400-milhoes-para-32-bilhoes-revela-onu/>. Acesso em: 15 fev. 2016.

[2] EXAME. *Mais da metade dos brasileiros são usuários da internet*. 2014. Disponível em: http://exame.abril.com.br/tecnologia/noticias/mais-da-metade-dos-brasileiros-sao-usuarios-da-internet. Disponível em: <http://exame.abril.com.br/tecnologia/noticias/mais-da-metade-dos-brasileiros-sao-usuarios-da-internet>. Acesso em: 15 fev. 2016.

[3] VIANNA, T. L. Hackers: um estudo criminológico da subcultura cyberpunk. *Revista do Centro Acadêmico Afonso Pena*, v. 6, n. 1, 2001.

[4] LEMOS, A. L. Ciber-rebeldes. *Universidad Federal de Bahia*, 1999. Disponível em: http://www. cfh. ufsc. br/˜ cso5421/bibliografias/rebelde. html.

[5] DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. 2. ed. New York: Wiley, 2001.

[6] TANASE, M. Ip spoofing: an introduction. *Security Focus*, v. 11, 2003.

[7] BREMLER-BARR, A.; LEVY, H. *Spoofing Prevention Method*. 2004.

[8] FERGUSON, P. Network ingress filtering: Defeating denial of service attacks which employ ip source address spoofing. 2000.

[9] BREMLER-BARR, A.; LEVY, H. Spoofing prevention method. In: IEEE. *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*. [S.l.], 2005. v. 1, p. 536–547.

[10] BIANCHI, M. F. de. *Extração de caracterÃsticas de imagens de faces humanas atravÃ©s de Wavelests, PCA e IMPCA*. Dissertação (Mestrado), 2006.

[11] LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007.

[12] GU, Q.; LI, Z.; HAN, J. Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725*, 2012.

[13] VAPNIK, V. N.; VAPNIK, V. *Statistical learning theory*. [S.l.]: Wiley New York, 1998. v. 1.

[14] CHAVES, A. d. C. F. *Extração de Regras Fuzzy para Máquinas de Vetores Suporte (SVM) para Classificação em Múltiplas Classes*. Tese (Doutorado) — PUC-Rio, 2006.

[15] SMOLA, A. J.; SCHÖLKOPF, B. *Learning with kernels*. [S.l.]: Citeseer, 1998.

[16] CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

[17] HORTA, E. G. Previsores para a eficiência da quimioterapia neoadjuvante no câncer de mama. *M. Sc., Universidade Federal de Minas Gerais (UFMG)*, 2008.

[18] SEMOLINI, R. *Support vector machines, inferência transdutiva e o problema de classificação*. Tese (Doutorado) — Universidade Estadual de Campinas, 2002.

[19] HORTA, E. G. et al. Extração de características e casamento de padrões aplicados à estimação de posição de um VANT. *UFMG*, 2011.

[20] LIMA, A. R. G. Máquinas de vetores suporte na classificaçao de impressoes digitais. *Universidade Federal do Ceará, Departamento de Computação, Fortaleza-Ceará*, 2002.

[21] JOACHIMS, T. *Learning to classify text using support vector machines: Methods, theory and algorithms*. [S.l.]: Kluwer Academic Publishers, 2002.

[22] KIM, K. I. et al. Support vector machines for texture classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 24, n. 11, p. 1542–1550, 2002.

[23] PONTIL, M.; VERRI, A. Support vector machines for 3d object recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 20, n. 6, p. 637–646, 1998.

[24] NOBLE, W. S. et al. Support vector machine applications in computational biology. *Kernel methods in computational biology*, p. 71–92, 2004.

[25] SCHOLKOPF, B.; GUYON, I.; WESTON, J. Statistical learning and kernel methods in bioinformatics. *Nato Science Series Sub Series III Computer and Systems Sciences*, IOS press, v. 183, p. 1–21, 2003.

[26] SILVA, R. et al. Segmentação, classificação e detecção de motociclistas sem capacete. *XI Simpósio Brasileiro de Automação Inteligente (SBAI), Fortaleza, Ceará–Brasil*, 2013.

[27] COSTA, J. A. L. d. et al. *Avaliação de dados de radar do sensor SAR-R99B no mapeamento do uso e cobertura da terra na Amazônia Central, município de Manaus, AM*. Dissertação (Mestrado), 2011.

[28] NASCIMENTO, R. F. F. et al. O algoritmo support vector machines (svm): avaliação da separação ótima de classes em imagens ccd-cbers-2. *Simpósio Brasileiro de Sensoriamento Remoto*, v. 14, p. 2079–2086, 2009.

[29] HUANG, C.; DAVIS, L.; TOWNSHEND, J. An assessment of support vector machines for land cover classification. *International Journal of remote sensing*, Taylor & Francis, v. 23, n. 4, p. 725–749, 2002.

[30] FOODY, G. M.; MATHUR, A. A relative evaluation of multiclass image classification by support vector machines. *Geoscience and Remote Sensing, IEEE Transactions on*, IEEE, v. 42, n. 6, p. 1335–1343, 2004.

[31] PAL, M.; MATHER, P. Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, Taylor & Francis, v. 26, n. 5, p. 1007–1011, 2005.

[32] SENNA, S. L. de. *Computação evolucionária Aplicada ao Diagnóstico de Falhas Incipientes em Transformadores de Potência Utilizando Dados de Cromatografia*. Dissertação (Mestrado), Setembro 2010.