Estudo Introdutório de Sazonalidades de Ataques a um IDS com o Uso do Algoritmo: KNN

Wellington Ferreira da Silva & Prof. ME. Eduardo Alves Moraes

Faculdade de Tecnologia de Ourinhos - FATECOU - Av. Vitalina Marcurso, 1400, Campus Universitário

wellington.silva112@fatec.sp.gov.br

Resumo

O avanço tecnológico das últimas décadas tem tornado indispensáveis alguns aparelhos e aparatos eletroeletrônicos no cotidiano das organizações. Atualmente, as empresas necessitam de redes de computadores sofisticadas, capazes de prover transmissão segura de informações. Por este motivo, elas buscam cada vez mais ferramentas de segurança da informação contra ataques. O IDS *Intrusion Detection System* - Sistema de detecção de intrusão. é uma ferramenta utilizada para combater ações maliciosas em ambientes de redes de computadores. Trabalham verificando padrões de comportamento nas redes e armazena essas informações em forma de *logs*, permitindo a análise posterior desses dados. Porém, devido ao grande volume de dados que está ferramenta gera, compreender essas informações se torna um desafio. Para auxiliar administradores de redes, neste artigo será apresentado o kNN *k Nearest Neighbour* - k-vizinhos mais próximos.

1. Introducão

Devido ao grande avanço tecnológico tomando como exemplo a Internet, em paralelo temos o desenvolvimento de várias técnicas de ataques que visam degradar ou interromper serviços fornecidos por empresas e organizações (4).

Sabendo-se que nenhum sistema pode ser totalmente seguro e que existem indivíduos interessados em obter informações sigilosas, um dos principais objetivos das organizações é manter seus dados seguros (3).

Uma ferramenta poderosa para detectar ataques a redes ou sistemas é o IDS (*Intrusion System Detection*), esse mecanismo quando detecta uma possível violação emite uma alerta e registra o arquivo em forma de *log* (3). Mecanismos de criptografia, antivírus, *firewalls*, políticas de segurança, algoritmos de aprendizagem de máquina como o kNN (*k Nearest Neighbour*), são exemplos dos muitos recursos que podem auxiliar na proteção de sistemas computacionais (2).

O desenvolvimento deste artigo busca auxiliar administradores de redes a compreender de maneira mais clara informações que podem ser extraídas de uma base dedados gerada por um IDS com auxílio de um algoritmo de aprendizagem de máquina.

2. Objetivos

Ilustrar o uso do algoritmo kNN em uma análise de uma base de dados gerada por um sistema de detecção de intrusão. Compõe os objetivos específicos estruturar o ambiente onde serão conduzidos os testes, analisar a base de dados com o auxílio do algoritmo kNN e apresentação dos resultados.

3. Metodologia

Para o desenvolvimento deste estudo, a base de dados utilizada foi obtida a partir de um IDS instalado em uma rede para coleta dos dados no período de uma semana. Após a coleta, essas informações foram organizadas em formato CSV para melhor compreensão. O algoritmo utilizado como principal ferramenta deste estudo foi o kNN e a medida utilizada para classificar a distância entre os k vizinhos foi a euclidiana. Para validação das classificações utilizou-se a matriz de confusão e acurácia, de acordo com (1) esse conceito exemplifica de forma clara os resultados de uma classificação e dispõe de avaliações singulares de cada categoria, bem como acertos e erros de exclusão.medida adotada para este estudo é a acurácia e pode ser definida como medida para identificar dados classificados como verdadeiros e falsos corretamente. Segue abaixo a fórmula desta medida:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

Para aplicação do algoritmo foram selecionadas quatro classes a serem avaliadas: regra de nome, prioridade, país e cidade. A cada classe foi aplicado um *script* dentro do ambiente de desenvolvimento da Linguagem de Programação R. O código foi escrito com um laço de repetição para que sejam efetuados 500 testes de forma que, a cada teste sejam utilizados novos modelos de dados para manter a especificidade de precisão das avaliações do algoritmo. Vale enfatizar que durante o experimento foram executados testes alterando o número de vizinhos k para calcular a acurácia e foi identificado que, mudar esta variável não tem impacto significativo no resultado final do cálculo, desta forma, para este artigo foi definido como padrão 13 k vizinhos a serem utilizados para classificação de novas instâncias.

Uma variável foi definida para alocar o resultado a cada repetição do *script*, em seguida todos os registros são somados e divididos pelo número total de testes obtendo-se a média geral de todas as avaliações do algoritmo.

4. Resultados e Discussão

Para aplicação do algoritmo foram selecionadas quatro classes a serem avaliadas: regra de nome, prioridade, país e cidade. Foi desenvolvido um código com um laço de repetição para que a cada teste sempre sejam utilizados novos modelos de dados para manter a especificidade de precisão das avaliações do algoritmo. Vale enfatizar que durante o experimento foram executados testes alterando o número de vizinhos k para calcular a acurácia e foi identificado que, mudar esta variável não tem impacto significante no resultado do cálculo. Foi definido como padrão 13 k vizinhos a serem utilizados para classificação de novas instâncias.abaixo segue uma tabela com os resultados obtidos:

Figure 1: *Resultado dos testes*

Classes	Média	Desvio Padrão	Coeficiente de Variação
Regra de Nome	53%	6,28	8,44%
Prioridade	88,75%	3,94	22.48%
País	36,52%	6,27	5.82%
Cidade	28,90%	5,92	4.87%
Media Geral	**51,79%**	**5,6**	**10,40%**

Fonte: *Elaboarado pelo autor*

5. Conclusão

Com bases nos resultados obtidos, observa-se que algoritmo retornou uma média geral de acurácia de previsão de acertos de 51,79%. É correto dizer que a escolha das classes a serem avaliadas implicou no resultado obtido, para um próximo estudo pode ser destacada a classe classificação, pois o IDS cria um grupo de classificação de alertas para os tipos de ataques identificados e pode ser interessante verificar o resultado. Este resultado já era esperado e isto afirma que o algoritmo kNN não é eficiente para tratar bases com grande volume de dados e, portanto, não é aconselhável aplicar em um ambiente real. Como sugestão para trabalhos futuros, pode-se fazer um estudo comparativo entre outros algoritmos que utilizam métricas diferentes de classificação como o *K-means*, algoritmo não supervisionado que se baseia na similaridade entre instâncias, *Random Forest*, algoritmo que utiliza o método de árvore decisão aleatória a fim de elevar o nível de acurácia, *Nave Baye*, um algoritmo classificador probabilístico, MPL 10 *(Multi-layer Perception)*, cria uma rede artificial de neurônios combinados em camadas para classificação de dados e SVM 11 *(Suport Vector Machine)*, algoritmo não supervisionado que cria hiperplanos que são utilizados para classificar as instâncias.

References

[1] ARRIGONI, T. R. Reconhecimento de silhueta de automóveis para carros autônomosutilizando aprendizado de máquina. 2018. *Citado na página 8.*

[2] MAINARDES, H. d. S. Análise e simulação de firewall. Dissertação (B.S. thesis) —Universidade Tecnológica Federal do Paraná, 2016. *Citado 2 vezes nas páginas 2 e 4.*

[3] MENEZES, F. R. L. de. *Comparação entre duas arquiteturas de redes sobre o ponto de vista de segurança: Um estudo de caso. 2011. Citado 2 vezes nas páginas 2 e 6.*

[4] MURINI, C. T. Análise dos sistemas de detecção de intrusão em redes: Snort e suricatacomparando com dados da darpa.Universidade Federal de Santa Maria, 2014. Citado 2 vezes nas páginas 2 e 5.

[5] ROBLES, W. T. de A. A. F. UTILIZAÇÃO DO ALGORITMO ÁRVORE DE DECISÃONA ANÁLISE DE ALERTA DE INTRUSÃO. Monografia — Faculdade de Tecnologia deOurinhos, 2017. *Citado 2 vezes nas páginas 2 e 3.*