Uso da Informação Mútua Ajustada na Seleção de Atributos numa Base de Dados de Detecção de Intrusos

Luiz E. R. Martins[1], **Nelcileno Virgílio de Souza Araújo**[1], **Allan G. de Oliveira**[1], **Letízia Manuella Serqueira Eugênio**[1]

[1]Instituto de Computação – Universidade Federal do Mato Grosso (UFMT)
Cuiabá – MT – Brasil

{luiz.martins,letizia.eugenio}@sou.ufmt.br

{nelcileno,allan}@ic.ufmt.br

Abstract. *Intrusion Detection Systems (IDSs) are essential for monitoring networks and identifying anomalous behavior. This paper applies a feature selection technique to extract the most representative characteristics from the NSL-KDD dataset, using a hybrid approach that combines the Information Gain Ratio and the K-means algorithm. The Adjusted Mutual Information (AMI) metric was employed to define the optimal subset of attributes. With this technique, it was possible to reduce the dimensionality from 41 to 7 attributes, achieving 70% accuracy, demonstrating the effectiveness of the proposed approach.*

Resumo. *Sistemas de Detecção de Intrusão (IDSs) são fundamentais para monitorar redes e identificar comportamentos anômalos. Este artigo aplica uma técnica de seleção de atributos para extrair as características mais representativas da base de dados NSL-KDD, utilizando uma abordagem híbrida que combina a Taxa de Ganho de Informação e o algoritmo K-means. A métrica de Informação Mútua Ajustada (IMA) foi empregada para definir o subconjunto ótimo de atributos. Com essa técnica, foi possível reduzir a dimensionalidade dos atributos de 41 para 7, alcançando uma acurácia de 70%, o que demonstra a eficácia da abordagem proposta.*

1. Introdução

Com o aumento exponencial dos incidentes de segurança, tornou-se essencial desenvolver sistemas de detecção de intrusões (IDSs) eficientes, que monitoram e identificam atividades anômalas nas redes [Liu and Lang 2019]. O conjunto de dados NSL-KDD, é amplamente utilizado para treinar e testar IDSs devido à sua estrutura mais equilibrada e representativa. O NSL-KDD possui 41 atributos, mas nem todos são igualmente relevantes para a detecção de intrusões, o que motiva abordagens que otimizem a seleção de características, reduzindo a dimensionalidade dos dados sem comprometer a eficiência dos sistemas.

Neste estudo, utilizamos a métrica de Informação Mútua Ajustada (IMA) como ferramenta de avaliação de modelos não supervisionados, visando melhorar a eficácia na seleção de atributos na base NSL-KDD. A metodologia utiliza uma técnica de seleção de atributos para extrair as características mais representativas, aplicando uma abordagem híbrida [Kayacik et al. 2005], [Lippmann et al. 2000]. Na primeira fase, usa-se a Taxa de

Ganho de Informação para medir a relevância dos atributos, enquanto na segunda fase, o algoritmo K-means é empregado para validar os agrupamentos com base nesses atributos selecionados. A métrica de Informação Mútua Ajustada (IMA) é utilizada para avaliar e refinar a seleção, garantindo que os atributos escolhidos representem de forma eficaz os padrões relevantes para o IDS [Kurniabudi et al. 2020].

2. Metodologia

Este estudo aplica a métrica IMA para medir a similaridade entre agrupamentos realizados, onde são comparados os rótulos reais e os preditos por técnicas de *Machine Learning*, além disso, ajuda a ajustar para casos de agrupamentos aleatórios. O cálculo que avalia a qualidade dos agrupamentos gerados pelo modelo de aprendizado não supervisionado é calculada a partir da comparação entre os rótulos reais dos dados e os rótulos preditos gerados pelo modelo. Na fórmula da IMA, representamos os rótulos reais por U e os rótulos preditos pelos agrupamentos gerados pelo modelo como V. A métrica calcula a Informação Mútua (IM) entre esses dois conjuntos de rótulos, que indica o quanto de informação os agrupamentos reais e preditos compartilham, refletindo a similaridade entre eles.

Além disso, a fórmula considera a expectativa de IM para agrupamentos aleatórios, indicada por $E[\text{IM}(U, V)]$, ajustando o valor da IMA para desconsiderar agrupamentos que ocorrem apenas por acaso. Também utilizamos as entropias de U e V, representadas por $H(U)$ e $H(V)$, que expressam a diversidade interna dos agrupamentos reais e preditos, respectivamente [Alessia Amelio 2016]. A normalização da fórmula, feita pela média das entropias, resulta em um valor final da IMA que varia de zero a um: valores próximos a um indicam uma forte correspondência entre os agrupamentos reais e preditos, enquanto valores próximos a zero refletem agrupamentos aleatórios. A sua fórmula é exibida abaixo.

$$\text{IMA}(U, V) = \frac{\text{IM}(U, V) - E[\text{IM}(U, V)]}{\frac{1}{2}(H(U) + H(V)) - E[\text{IM}(U, V)]}$$

Legenda das váriaveis:

- U: rótulos reais.
- V: rótulos preditos.
- $\text{IM}(U, V)$: Informação Mútua
- $E[\text{IM}(U, V)]$: expectativa da Informação Mútua.
- $H(U)$ e $H(V)$: entropias.

A escolha da IMA justifica-se por sua eficiência em técnicas de aprendizado não supervisionado, onde métricas mais clássicas, como a acurácia, apresentam limitações. A acurácia, por exemplo, requer rótulos conhecidos para cada instância, o que não está disponível em cenários de aprendizado não supervisionado. Além disso, a acurácia não é sensível à qualidade das divisões internas dos grupos, sendo necessário outras métricas para serem usadas para análise. Assim, a IMA é mais adequada, pois permite uma avaliação mais robusta ao considerar a probabilidade de agrupamentos aleatórios em técnicas não supervisionadas [Lazarenko and Bonald 2021].

A base de dados NSL-KDD foi utilizada nesta pesquisa por ser amplamente empregada na avaliação e desenvolvimento de sistemas de detecção de intrusões (IDSs)

[Kayacik et al. 2005], [Lippmann et al. 2000]. Esse conjunto de dados é composto por conexões de rede classificadas em categorias que incluem tráfego normal e tipos variados de ataques, como negação de serviço (DoS), varreduras (Probe), tentativas de acesso local remoto (R2L) e elevações de privilégio de usuário para administrador (U2R) [Kayacik et al. 2005], [Lippmann et al. 2000]. Cada conexão é descrita por 41 atributos que capturam informações sobre o tráfego de rede, incluindo características básicas do protocolo, dados de conteúdo e propriedades temporais, o que possibilita uma análise abrangente dos padrões de comportamento da rede [Kayacik et al. 2005], [Lippmann et al. 2000]. Sendo assim, este estudo visa selecionar o melhor subconjunto de características da base de dados NSL-KDD com o suporte de uma métrica eficaz.

A metodologia segue duas etapas principais. Na primeira, utiliza-se a Taxa de Ganho de Informação, que quantifica a relevância de cada atributo, para selecionar características no estilo da técnica filter. Já na segunda etapa, aplica-se o K-means na técnica wrapper para validar e agrupar os atributos mais relevantes utilizando a técnica de validação *10-fold cross-validation* onde o conjunto de dados é dividido em dez partes iguais. Em cada rodada, nove partes são usadas para treinar o modelo e uma para teste, repetindo o processo dez vezes para que cada parte seja usada para teste uma vez. No final, a média dos resultados de todas as rodadas fornece uma avaliação mais confiável do desempenho do modelo, reduzindo variância e generalização [Zhang and Liu 2023].

Esta abordagem híbrida foi inspirada na metodologia de [Araújo et al. 2010], escolhida pela sua eficácia em otimizar a seleção de atributos, mas adaptada neste estudo para incluir a métrica IMA.

3. Resultados e Discussão

Avaliar o desempenho de diferentes métricas em modelos de aprendizado não supervisionado é fundamental para garantir uma boa representação na etapa de validação. A métrica IMA, mostra-se eficaz ao comparar o agrupamento gerado com o agrupamento verdadeiro, considerando também a possibilidade de agrupamentos aleatórios.

A IMA, como variação da métrica de Informação Mútua, ajusta a avaliação dos agrupamentos ao considerar a probabilidade de agrupamentos aleatórios [Lazarenko and Bonald 2021]. Utilizando esta métrica, foi possível identificar que o melhor subconjunto de características contém sete atributos.

Com o melhor subconjunto de sete atributos, foi realizada a validação cruzada nos dez subconjuntos da base KDD99, resultando em uma média de acurácia de 70,80%. Comparando com a abordagem de [Araújo et al. 2010], que utilizou 14 atributos, nossa pesquisa obteve uma redução de 50% no número de atributos necessários para alcançar resultados satisfatórios, o que demonstra a eficiência da metodologia proposta em termos de redução de dimensionalidade e desempenho como mostrado na Tabela 1.

Metodologia	Métrica	Base de Dados	Atributos / Redução	Desempenho
[Araújo et al. 2010]	Acurácia	NSL-KDD	14 atributos / 65%	99%
Nossa Proposta	IMA	NSL-KDD	7 atributos / 80%	70%

Tabela 1. Comparação entre Acurácia e Informação Mútua Ajustada na base de dados NSL-KDD

4. Conclusão

Neste artigo, foi apresentada uma nova abordagem para a avaliação de modelos não supervisionados, utilizando a métrica de Informação Mútua Ajustada (IMA). Esta métrica se mostrou eficiente para medir a similaridade entre clusters, destacando-se por sua simplicidade e objetividade. Sua aplicação no contexto deste estudo evidenciou a viabilidade de seu uso em problemas de agrupamento, especialmente por se ajustar bem à natureza não supervisionada das tarefas.

Para fornecer uma comparação concreta, revisamos o estudo de [Araújo et al. 2010], que utilizou métricas clássicas da literatura. Em seu trabalho, foi necessário consumir aproximadamente 35% dos atributos para alcançar resultados satisfatórios, enquanto, com a abordagem proposta neste artigo, foi possível obter resultados equivalentes utilizando apenas cerca de 17% dos atributos, o que representa uma redução significativa na quantidade de dados processados.

Para trabalhos futuros, pretende-se explorar a avaliação da matriz de confusão em classificações entre normal e anômalo, além de investigar métricas como acurácia, taxa de detecção global e índice kappa. Outra linha de pesquisa será testar a abordagem híbrida com técnicas adicionais de aprendizado de máquina para aprimorar a seleção e a avaliação de atributos.

Referências

Alessia Amelio, C. P. (2016). Correction for closeness: Adjusting normalized mutual information measure for clustering comparison.

Araújo, N., de Oliveira, R., Ferreira, E., Shinoda, A. A., and Bhargava, B. (2010). Identifying important characteristics in the kdd99 intrusion detection dataset by feature selection using a hybrid approach. In *2010 17th International Conference on Telecommunications*, pages 552–558.

Kayacik, H. G., Zincir-Heywood, A. N., and Heywood, M. I. (2005). Selecting features for intrusion detection: A feature relevance analysis on kdd 99 intrusion detection datasets. In *Proceedings of the third annual conference on privacy, security and trust*, volume 94, pages 1723–1722. Citeseer.

Kurniabudi, K., Stiawan, D., Dr, D., Idris, M., Bamhdi, A., and Budiarto, R. (2020). Cicids-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access*, PP:1–1.

Lazarenko, D. and Bonald, T. (2021). Pairwise adjusted mutual information. *CoRR*, abs/2103.12641.

Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., and Das, K. (2000). The 1999 darpa off-line intrusion detection evaluation. *Computer networks*, 34(4):579–595.

Liu, H. and Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences*, 9(20).

Zhang, X. and Liu, C.-A. (2023). Model averaging prediction by k-fold cross-validation. *Journal of Econometrics*, 235(1):280–301.