Problemas e alternativas ao uso da base KDDCUP'99, para classificação de ataques em redes IP

Ricardo Balbinot[1,2], **Alexandre Balbinot**[2], **Ivan Müller**[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS)
Campus Canoas

[2]Programa de Pós-graduação em Engenharia Elétrica
Universidade Federal do Rio Grande do Sul (UFRGS)

ricardo.balbinot@canoas.ifrs.edu.br

alexandre.balbinot@ufrgs.br, ivan.muller@ufrgs.br

Abstract. *This document aims to address the limitations of the KDDCUP'99 dataset, suggesting the UNSW-NB15 dataset as an alternative. To examine the main issue of the KDD dataset – the ease of data classification – three classification methods are used: random forests, SVM, and neural networks.*

Resumo. *O presente documento busca abordar as limitações da base KDDCUP'99, sugerindo como alternativa de uso da base UNSW-NB15. Para verificar o principal problema da base KDD, a facilidade de classificação dos dados, são utilizados três métodos de classificação: florestas aleatórias, SVM e redes neurais.*[1]

1. Introdução

A detecção e reação à tentativas de invasões em uma rede é uma atividade essencial. Um ponto central desses sistemas reside nas classificação dos fluxos da rede como normais ou que representem algum ataque.

A base KDDCUP'99 é uma base amplamente utilizada nesse contexto, sendo considerada inclusive como um parâmetro de comparação de desempenho na proposição de novos métodos. Contudo, os diversos problemas desse *dataset* devem ser considerados, com o uso da base sendo, atualmente, desaconselhado. A base apresenta limitações diversas, que vão desde a presença de registros duplicados, desbalanceamento dos dados (vide [Tavallaee et al. 2009, Sapre et al. 2019]), e até mesmo considerações sobre a representatividade da base em termos dos fluxos realmente observados [Moustafa e Slay 2015].

O presente estudo busca apresentar e discutir algumas das limitações da base em questão, propondo, em sua substituição, o uso do *dataset* UNSW-NB15 [Moustafa e Slay 2015]. Para indicar alguns dos problemas observados na utilização da KDDCUP e o uso da UNSW-NB15 como alternativa, uma análise de classificação binária dos fluxos de ambas é realizada com as técnicas de florestas aleatórias, máquinas de vetor de suporte e redes neurais artificiais.

[1]O presente trabalho foi realizado com apoio do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS).

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil(CAPES) - Código de Financiamento 001.

2. Pesquisa bibliográfica

Tavallaee et al. [Tavallaee et al. 2009] propõem uma análise detalhada da KDDCUP'99, indicando severas limitações da mesma. Um dos pontos críticos refere-se à facilidade de classificação, visto que 98% dos registros de treino e 86% de teste foram corretamente classificados por todos os modelos. Os autores propuseram uma nova versão da base: a NSL-KDD, refletindo melhor o desempenho real de diferentes abordagens.

Moustafa e Slay [Moustafa e Slay 2015] descrevem a base UNSW-NB15. No trabalho a base é indicada como evolução face à KDDCUP e NSL-KDD: possui maior diversidade de ataques, maior número de atributos e captura em múltiplas redes.

Adhikary et al. [Adhikary et al. 2021] realizam uma análise do uso de *kernels* SVM (*Support Vector Machines*) para detecção de ataques, com a KDDCUP'99. Os autores indicam que a função Laplaciana teve o melhor desempenho geral.

Sapre et al. [Sapre et al. 2019] fazem uma análise comparativa entre a KDDCUP e NSL-KDD, com o uso das técnicas: Naïve Bayes, SVM , Florestas Aleatórias e Redes Neurais Artificiais (RNA). O objetivo principal é avaliar o impacto da qualidade das bases de dados no desempenho dos modelos.

Meliboev et al. [Meliboev et al. 2022] avaliam a eficácia de modelos para detecção de intrusões, usando diferentes RNAs. O objetivo é comparar o desempenho com dados balanceados e desbalanceados, utilizando: KDDCUP'99, NSL-KDD e UNSW-NB15.

3. Metodologia

Para a comparação entre as bases KDDCUP e UNSW foi feito o treinamento e uso de florestas aleatórias, SVM, e RNAs, para a classificação binária dos fluxos.

Após análise exploratória da base KDD, foi necessário proceder ao seu tratamento, com passos representando alguns problemas observados na mesma. Dentre as etapas estão a eliminação de duplicatas (na base completa são 78% dos registros), balanceamento e normalização. A base é fortemente desbalanceada, com 75,6% apontando para um fluxo normal e os restantes 24,5% para ataques. O balanceamento foi feito via *undersampling*.

Os testes consideraram dois casos: o uso de florestas aleatórias e SVM, e o uso de RNAs. No primeiro caso, os testes consideraram duas situações: uso da base completa, tratada, e uso da base reduzida, com as 10 principais características indicadas pela permutação de florestas aleatórias. Para o segundo caso, foram três modelos utilizados: base completa, não tratada, base completa, tratada, e base reduzida, com as mesmas 10 características do caso anterior. Os mesmos testes foram conduzidos com a base UNSW.

A base UNSW pode ser trabalhada na sua forma completa, ou nas versões específicas de testes e treinos. A base completa possui 2540044 registros de fluxos, com a versão de treino e testes tendo, respectivamente, 175341 e 82332 registros. Para as avaliações realizadas, foram usadas as versões de treino e testes. A base de treino da UNSW-NB15 é fortemente desbalanceada. No seu tratamento foi feito o balanceamento, via *oversampling* com SMOTE, e normalização dos dados.

4. Resultados e discussões

Iniciando a discussão pelo comportamento observado na base KDDCUP, o primeiro teste – resultados apresentados da tabela 2 – considerou a classificação da base original, com

todas as características. Foi utilizada a base completa, com 4898430 registros. As principais características dos modelos de classificação utilizados podem ser vistas na tabela 1. Em seguida, foram conduzidos os testes com a base tratada e com a base reduzida.

Tabela 1. Principais características dos modelos utilizados

Base	Modelo	Características
KDDCUP'99	RNA (base completa)	Duas camadas ocultas, 41 e 20 neurônios, ativação sigmóide, 30 épocas de treino, parada precoce
	RNA (base reduzida)	Idem ao caso acima, com 50 épocas, parada precoce
	RF	200 estimadores, critério de Gini
	SVM	Kernel sigmóide, $C = 1.0$
UNSW-NB15	RNA (base completa)	Duas camadas ocultas, 30 e 15 neurônios, ativação sigmóide, 50 épocas, parada precoce
	RNA (base reduzida)	Duas camadas ocultas, 200 e 45 neurônios, ativação sigmóide e relu, *dropout* de 45% e 35%, taxa de aprendizagem 0.01, 50 épocas, parada precoce
	RF	200 estimadores, critério de Gini
	SVM	Kernel sigmóide, $C = 1.0$

A expectativa era de uma situação com um baixo índice de acerto, particularmente no caso da base completa não tratada. Contudo, foi percebido que, em quase todos os cenários, os índices de classificação correta foram superiores a 99%. Ou seja, a base realmente aparenta ter uma taxa significativamente elevada de classificação, independentemente do método utilizado, o que, por si só, já desaconselha o seu uso. Além disso, cabe destacar que a base, pós eliminação de duplicatas, apresenta uma má distribuição de ataques, o que compromete a sua utilidade.

Por outro lado, podemos ver os resultados, na tabela 2, da classificação, usando os mesmos métodos, para a base UNSW-NB15, onde fica evidente que o uso de diferentes técnicas e o devido tratamento dos dados gera resultados significativamente distintos. A base UNSW oferece um conjunto de dados também com uma melhor representação dos diferentes ataques e sem um comportamento de viés.

5. Conclusões

A partir das discussões efetuadas na seção anterior, resta claro que o uso de classificadores pelo método de florestas aleatórias ainda se mostra bastante interessante. Contudo, como ponto mais significativo, faz-se necessário considerar os problemas no uso da base KDDCUP'99 – particularmente no que tange o alto índice de classificações com sucesso independentemente da técnica utilizada. Como vários autores indicam ([Tavallaee et al. 2009] e [Moustafa e Slay 2016]), é necessário abandonar o uso dessa base em prol de bases mais recentes e melhor estruturadas. A base UNSW-NB15 [Moustafa e Slay 2015] se mostra bem interessante, muito embora ainda apresente um desbalanceamento elevado.

Como sugestão de continuidade para este trabalho, pode-se realizar a busca e análise de outras bases, mais recentes, capazes de representar melhor as características

Tabela 2. Resultados de classificação nos diferentes cenários

Base	Caso	Técnica	*Accuracy*
KDDCUP'99	Base completa, não tratada	RNA	99,9818%
	Base completa, tratada	RNA	99,9389%
		SVM	91,4670%
		RF	99,9676%
	Base reduzida, tratada	RNA	99,6977%
		SVM	96,2704%
		RF	99,9530%
UNSW-NB15	Base completa, não tratada	RNA	83,4013%
	Base completa, tratada	RNA	87,6342%
		SVM	75,1149%
		RF	96,8871%
	Base reduzida, tratada	RNA	92,5173%
		SVM	64,2022%
		RF	96,6189%

de ataques nas redes atuais – de fato, mesmo a UNSW-NB15 já representa um retrato de 10 anos atrás; os ataques nesse período se modificaram significativamente.

Referências

Adhikary, K., Bhushan, S., Kumar, S., e Dutta, K. (2021). Evaluating the performance of various svm kernel functions based on basic features extracted from kddcup'99 dataset by random forest method for detecting ddos attacks. *Wireless Personal Communications*, 123(4):3127–3145.

Meliboev, A., Alikhanov, J., e Kim, W. (2022). Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets. *Electronics*, 11(4):515.

Moustafa, N. e Slay, J. (2015). Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). Em *2015 Military Communications and Information Systems Conference (MilCIS)*, páginas 1–6. IEEE.

Moustafa, N. e Slay, J. (2016). The evaluation of network anomaly detection systems: Statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set. *Information Security Journal: A Global Perspective*, 25(1–3):18–31.

Sapre, S., Ahmadi, P., e Islam, K. (2019). A robust comparison of the kddcup99 and nsl-kdd iot network intrusion detection datasets through various machine learning algorithms.

Tavallaee, M., Bagheri, E., Lu, W., e Ghorbani, A. A. (2009). A detailed analysis of the kdd cup 99 data set. Em *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, páginas 1–6. IEEE.