Análise do impacto de viés nos conjuntos de dados para detecção de Malwares Android

Lucas Vilanova¹, Renato Sayyed¹, Taina Soares¹,
Guilherme Siqueira¹, Gustavo Rodrigues¹, Eduardo Feitosa², Diego Kreutz¹

¹ Universidade Federal do Pampa (Unipampa)

² Universidade Federal do Amazonas (UFAM)

{NomeSobrenome}.aluno@unipampa.edu.br
efeitosa@icomp.ufam.br, kreutz@unipampa.edu.br

Resumo. Atualmente, a detecção de malwares Android é realizada, majoritariamente, através de modelos de aprendizado de máquina. O problema é que a maioria dos modelos desenvolvidos têm sido treinados com conjuntos de dados defasados (e.g., de 2012). Nosso objetivo é coletar evidências iniciais para demonstrar o impacto de diferentes datasets no desempenho de modelos preditivos. Para isto, utilizamos conjuntos de dados de diferentes períodos temporais, isto é, de 2012 a 2021.

1. Introdução

O sistema Android é atualmente um alvo frequente de aplicações maliciosas (*malwares*). Estudos apontam que modelos de aprendizagem de máquina (*machine learning*) conseguem detectar *malwares* com uma acurácia acima de 90% [Sharma and Rattan, 2021]. Contudo, os *malwares* aumentam em número e sofisticação, exigindo atualização e desenvolvimento de novos modelos [Sahay et al., 2020].

Como modelos de *machine learning* aprendem com os dados de entrada, o desempenho (i.e., capacidade preditiva) destes é impactado fortemente pelos *datasets* (conjunto de dados) utilizados para treino [Allix et al., 2015]. Ainda assim, a maioria das soluções aplicadas à detecção de *malwares* utiliza *datasets* antigos (e.g., The Drebin Dataset, de 2012) ou defasados (e.g., características de APIs antigas) [Soares et al., 2021b].

Levantamos a hipótese de que ao treinarmos modelos de *machine learning* com um *dataset* defasado, estes não irão apresentar o mesmo poder preditivo para conjuntos de dados atuais. Dentre os diversos indícios que corroboram a hipótese, podemos destacar: a evolução das APIs dos aplicativos Android (da versão 1 em 2008 a versão 31 em 2021); a constante evolução e sofisticação do comportamento dos aplicativos maliciosos; e o aumento expressivo na quantidade de *malwares*.

Com o objetivo de verificar a hipótese levantada, neste trabalho propomos um experimento para a análise do impacto no desempenho de um modelo de aprendizagem de máquina, a partir de conjuntos de dados de diferentes períodos de tempo (i.e., mais antigos e mais atuais). Para a construção do modelo, escolhemos o algoritmo Random Forest [Breiman, 2001] por ser um dos mais utilizados na detecção de *malwares* Android [Sharma and Rattan, 2021].

As contribuições do trabalho podem ser resumidas em: (a) caracterização do impacto da utilização de uma mesma configuração de um modelo, adequado a um conjunto

de dados defasados, em *datasets* recentes; e (b) identificação de dados duplicados em *datasets* amplamente utilizados na literatura, bem como demonstração dos efeitos desses dados nos modelos de aprendizado de máquina.

O trabalho está organizado como segue. Nas Seções 2, 3 e 4 apresentamos os *datasets*, o modelo e os resultados, respectivamente. Na Seção 5 discutimos a atualidade dos *dataset* e, finalmente, na Seção 6 apresentamos as considerações finais.

2. Datasets

Os 4 (quatro) *datasets* utilizados, selecionados a partir do levantamento apresentado em [Soares et al., 2021b], são apresentados na Tabela 1. Para a seleção dos *datasets*, utilizamos dois critérios: (1) o mais popular na literatura, naquele ano, e (2) os que estavam publicamente disponíveis. Selecionamos então, conjuntos de dados com o intervalo de 3 anos (2012, 2015, 2018 e 2021).

Tabela 1. Datasets

Datasets	Data	Com Duplicatas		Sem Duplicatas		Permissões	Permissões após limpeza
		Benignos	Malignos	Benignos	Malignos		
Drebin-215	2012	9.476	5.560	5.539	1.720	114	113
Androcrawl	2014-2015	139.670	23.313	138.867	23.247	61	58
AndroidMalwareNormal	2018	18.850	9.999	16.698	7.026	173	173
DefenseDroid	2021	5.975	6000	3.146	4.579	1.490	134

Conjuntos de dados como o o Drebin-215 e Androcrawl contêm características como permissões, chamadas de API e intenções. Entretanto, como o objetivo é utilizar apenas permissões, pois são as características mais relevantes e utilizadas [Sharma and Rattan, 2021], removemos as demais características. Ao final do processo de limpeza e seleção de apenas permissões, os *datasets* Drebin-215, Androcrawl, AndroidMalwareNormal e DefenseDroid ficaram com 113, 58, 173 e 134 permissões, respectivamente.

A preparação dos dados é um conjunto de etapas fundamentais para melhorar e adequar o conjunto de dados ao algoritmo de aprendizado de máquina [Zheng and Casari, 2018]. Por exemplo, a preparação pode evitar ruídos nos dados (e.g., como no caso do Androcrawl *dataset*, que continha todos os valores com formato textual e apresentava valores inválidos) que poderiam causar erros nos cálculos do modelo por não estarem em um formato numérico ou por apresentarem valores faltantes [Qi et al., 2018]. A preparação dos 4 *datasets* incluiu a exclusão de características (e.g., a permissão SET_WALLPAPER, duplicada no Drebin-215) de forma a auxiliar o modelo a realizar seus cálculos.

No *dataset* Androcrawl identificamos três permissões com apenas valores “0” (RECEIVE, MESSAGE, SEND) e com 10.262 exemplos de aplicativos Android com valores inválidos (i.e., “?”). Para que o *dataset* não sofresse uma redução significativa nos

exemplos de aplicativos caso removêssemos cada registro com valores faltantes, essas três permissões foram removidas.

No DefenseDroid removemos um conjunto de 1.356 *features*, incluindo duplicatas e permissões que não fazem parte da lista de permissões do Android¹, resultando em um *dataset* de 134 permissões. No caso do AndroidMalwareNormal, foram removidas apenas as características “Category” e “Package”, que não são permissões.

O que mais nos chamou atenção na etapa de preparação dos dados foram os exemplos de aplicativos Android duplicados. Há 7.777, 5.125, 4.250 e 869 duplicadas nos *datasets* Drebin-215, AndroidMalwareNormal, DefenseDroid e Androcrawl, respectivamente. Os exemplos duplicados em *datasets* podem impactar no desempenho dos algoritmos de aprendizado de máquina, podendo levar ao *overfitting* e enviesamento do modelo [Zhao et al., 2021]. Removemos também todas as duplicatas, mantendo apenas a primeira ocorrência.

Os dados originais, que passaram pelo processo de limpeza e correção, bem como os conjuntos de dados resultantes e suas análises exploratórias, utilizados neste trabalho, estão disponíveis em <https://github.com/Malware-Hunter/Dataset-Matters>.

3. Modelo RandomForest

O algoritmo RandomForest é baseado em diversas árvores de decisão utilizadas em amostras de dados selecionadas aleatoriamente. A partir das previsões de cada árvore é selecionada a melhor solução por meio de votação.

Os dados de cada um dos 4 *datasets* foram separados em três partes, sendo 40% para treino, 30% para validação e 30% para teste. Além disso, devido à diferença da amostragem dentre *datasets*, resolvemos padronizar para o menor conjunto. Isto reduz a probabilidade de que um *dataset* aprenda mais que outro por conta da quantidade de amostras de aplicativos. Para isso, selecionamos a menor quantidade de cada classe entre todos os *datasets* (3.146 benignos e 1.720 malignos) e a padronizamos como quantidade total de exemplos fornecidos ao conjunto de treino nos 4 conjuntos de dados. Assim, o impacto da diferença da quantidade de exemplos treinados em cada *dataset* no desempenho do modelo é anulado. Utilizamos o parâmetro `random_state=42` para selecionar os mesmos conjuntos de dados em cada partição e não modificar os resultados a cada treino.

Como o nosso objetivo é classificar aplicativos entre benignos e malignos, considerando conjuntos de dados desbalanceados, utilizamos a métrica ROC-AUC (ou curva ROC). A curva ROC é uma métrica de avaliação que descreve a capacidade do modelo de distinguir entre duas classes (benignos e malignos), onde é necessário calcular as probabilidades de cada observação (exemplo de um aplicativo) pertencer a uma classe. O valor de ROC-AUC varia de 0,0 até 1,0, onde o 0,0 representa um modelo que erra todas as previsões e 1,0 indica um teste perfeito de predição.

¹<https://developer.android.com/reference/android/Manifest.permission>

3.1. Otimização do modelo

O experimento consiste em analisar o desempenho de um modelo, configurado para um *dataset* defasado, em conjuntos de dados mais recentes. O *dataset* de referência selecionado foi o Drebin-215 (de 2012).

Primeiramente, foi necessário balancearmos os dados de treino a fim de nivelar a proporção entre exemplos de aplicações benignas e malignas, pois como a classe majoritária (benignos) está presente em excesso, isso pode ocasionar em *overfitting* no algoritmo, fazendo com que o modelo acabe errando muitos exemplos da classe minoritária (malignos).

Com o conjunto de dados de treino balanceado, treinamos o classificador Random-Forest com seus parâmetros *default* (i.e., parâmetros padrão fornecidos pela biblioteca de *scikit-learn*²) e o validamos com dados desbalanceados para aproximar o teste de um caso real. Com isto, obtivemos um resultado de 89,47% de ROC-AUC, e com uma garantia maior de não incluir viés no modelo.

O passo seguinte no desenvolvimento de um modelo é ajustar os hiperparâmetros do algoritmo, pois eles impactam nas taxas de detecção [Probst et al., 2019]. Utilizamos a função `RandomizedSearchCV` da biblioteca `scikit-learn`³. O espaço de busca dos hiperparâmetros, bem como a combinação ótima de valores encontrada para esse experimento estão descritos na Tabela 2. O resultado do modelo otimizado no *dataset baseline* Drebin-215, utilizando o treino balanceado, foi de 89,58% de ROC-AUC.

Tabela 2. Hiperparâmetros do modelo RandomForest

Hiperparâmetro	Padrão	Intervalo de busca	RandomizedSearchCV
n_estimators	100	200 : 2000	1.800
min_samples_split	2	2, 5 e 10	2
min_samples_leaf	1	1, 2 e 4	1
max_features	auto	auto e sqrt	auto
max_depth	None	10 : 110 e None	20
bootstrap	True	True e False	False

4. Resultados

A partir dos resultados apresentados na Tabela 3, podemos concluir que o modelo configurado para o Drebin-215 obteve resultados muito baixos nos *datasets* Androcrawl e AndroidMalwareNormal. Isso se deve ao fato desses conjuntos de dados conterem características não relevantes aos cálculos do algoritmo, prejudicando a capacidade preditiva do modelo. Isto, por si só, já corrobora parcialmente a nossa hipótese: é pouco efetivo treinar um modelo com *datasets* defasados e aplicá-lo, posteriormente, a *datasets* atuais. O *dataset* impacta de forma significativa no desempenho do modelo.

Além disso, podemos observar que as duplicatas também podem causar impacto no desempenho do algoritmo. Elas podem levar o modelo ao enviesamento, pois ao ser

²<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

³https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Tabela 3. Resultados do modelo RandomForest

Dataset	ROC-AUC	
	Com duplicata	Sem duplicata
Drebin-215 (2012)	94,37%	89,58%
Androcrawl (2014/2015)	55,80%	54,39%
AndroidMalwareNormal (2018)	55,80%	50,47%
DefenseDroid (2021)	84,61%	84,85%

testado com um conjunto de dados que contenha exemplos de aplicativos que o mesmo já treinou, ele saberá prever corretamente [Zhao et al., 2021]. Portanto, eliminar duplicatas é fundamental para treinar um modelo de detecção de *malwares*.

Na prática, o desempenho do modelo decaiu em 4,79% para o *dataset* Drebin-215 com e sem duplicatas. Como o Drebin-215 continha mais de 50% de exemplos duplicados, o modelo poderia estar enviesado. Com a remoção das duplicatas, temos uma garantia maior de termos um modelo que não sofre de enviesamento.

As duplicatas no *dataset* AndroidMalwareNormal também impactaram o desempenho do modelo (em 5,33%). Assim, como no caso do Drebin-215, a redução no resultado, em parte, pode ter sido causada pelo enviesamento e pela redução significativa de exemplos de aplicativos. Já no caso do *dataset* DefenseDroid, apesar de possuir 35,49% de dados duplicados, o impacto no resultado do modelo foi insignificante. Isso ocorre pelo fato de uma boa parcela das permissões do *dataset* serem classificadas como relevantes para a detecção de *malwares* [Li et al., 2018, Assolin et al., 2021].

5. Discussão

Existe uma questão bastante peculiar com relação à atualidade dos dados dos *datasets*. Inicialmente assumimos que a data do *dataset* é a mesma da coleta dos dados. Entretanto, durante nossas análises, identificamos que existe uma probabilidade de que os dados sejam mais antigos do que o esperado (e.g., do que o informado no *dataset*). Esse problema ocorre quando os *datasets* possuem dados de versões de APIs muito antigas e não possuem nenhum dado de versões das APIs do Android (e.g., versão da API do mesmo ano de criação do *dataset* ou do ano anterior). Isto pode ocorrer por diferentes razões, como discutimos detalhadamente em outro trabalho [Soares et al., 2021a]. Por exemplo, é comum os autores compilarem *datasets* a partir de outros *datasets* previamente existentes (e mais antigos). Consequentemente, as características dos aplicativos (e.g., permissões, chamadas de API) acabam sendo defasadas.

6. Considerações Finais

Nesse artigo, avaliamos o impacto de *datasets* de diferentes datas no desempenho de um modelo baseado no algoritmo RandomForest. O experimento consistiu em treinar e configurar o modelo para o *dataset* Drebin-215 e avaliar seu desempenho em *datasets* mais recentes, utilizando todas as permissões presentes nos conjuntos de dados.

Os resultados obtidos indicam que há uma forte relação entre o *dataset* e o desempenho do modelo. Simplesmente utilizar um *dataset* defasado não significa que o modelo

irá performar bem com *datasets* mais atuais. Primeiro, os *datasets* podem conter características distintas, o que vai levar o modelo a resultados diferentes do esperado quando este for treinado com dados defasados. Segundo, diferentemente dos *datasets* defasados, *datasets* recentes podem incorporar conjuntos de características das novas versões das APIs do Android, que não existiam anteriormente, impactando o desempenho dos modelos. Portanto, para detectarmos *malwares* atuais, é fundamental o treinamento dos novos modelos com dados atuais.

Como trabalhos futuros, podemos destacar: (a) a criação de um *dataset* com dados atuais, pois os que existem na literatura não contemplam as últimas versões das APIs do Android e não incorporam amostras significativas de aplicativos maliciosos e benignos de 2020 e 2021; (b) investigar extensivamente os viesamentos causados pelos dados de *datasets* existentes (e.g., registros duplicatas); e (c) investigar questões relacionadas a rastreabilidade das amostras dos *datasets*.

Agradecimentos

Esta pesquisa foi financiada, conforme previsto nos Arts. 21 e 22 do decreto no. 10.521/2020, nos termos da Lei Federal no. 8.387/1991, através do convênio no. 003/2021, firmado entre ICOMP/UFAM, Flextronics da Amazônia Ltda e Motorola Mobility Comércio de Produtos Eletrônicos Ltda.

Referências

- Allix, K., Bissyandé, T. F., Klein, J., and Le Traon, Y. (2015). Are your training datasets yet relevant? In *International Symposium on Engineering Secure Software and Systems*, pages 51–67. Springer.
- Assolin, J., Rocha, V., Kreutz, D., Siqueira, G., Rodrigues, G., Feitosa, E., and Casola, K. (2021). Detecção de Malwares Android: reprodução da seleção de características do SigPID. https://arxiv.kreutz.xyz/wrseg2021_sigpid_vel.pdf.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Li, J., Sun, L., Yan, Q., Li, Z., Srisa-An, W., and Ye, H. (2018). Significant permission identification for machine-learning-based android malware detection. *IEEE Transactions on Industrial Informatics*, 14(7):3216–3225.
- Probst, P., Boulesteix, A.-L., and Bischl, B. (2019). Tunability: importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, 20(1):1934–1965.
- Qi, Z., Wang, H., Li, J., and Gao, H. (2018). Impacts of dirty data: and experimental evaluation. *arXiv preprint arXiv:1803.06071*.
- Sahay, S. K., Sharma, A., and Rathore, H. (2020). Evolution of malware and its detection techniques. In *Information and Communication Technology for Sustainable Development*, pages 139–150. Springer.
- Sharma, T. and Rattan, D. (2021). Malicious application detection in android—a systematic literature review. *Computer Science Review*, 40:100373.
- Soares, T., Assolin, J., Barcellos, L., Sayyed, R., Casola, K., Kreutz, D., Costa, E., Gustavo, N., and Feitosa, E. (2021a). Detecção de malwares android: Disponibilidade e atualização das fontes de dados. https://arxiv.kreutz.xyz/wrseg2021_disponibilidade_vel.pdf.
- Soares, T., Siqueira, G., Barcellos, L., Sayyed, R., Vargas, L., Rodrigues, G., Assolin, J., Pontes, J., and Kreutz, D. (2021b). Detecção de malwares android: datasets e reprodutibilidade. https://arxiv.kreutz.xyz/mh21_reprodutibilidade.pdf.
- Zhao, Y., Li, L., Wang, H., Cai, H., Bissyandé, T. F., Klein, J., and Grundy, J. (2021). On the impact of sample duplication in machine-learning-based android malware detection. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 30(3):1–38.
- Zheng, A. and Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. "O'Reilly Media, Inc."