Aplicabilidade e Impactos quanto a Adoção de Modelos de Classificação como Mecanismos Anti-phishing.

Mateus L. S. D. de Barros[1], Carlo M. R. da Silva[2], Péricles C. B. de Miranda[1]

[1]DEInfo – Universidade Federal Rural de Pernambuco (UFRPE)

[2]Universidade de Pernambuco (UPE)

mateuslins02@gmail.com, revoredo@gmail.com, periclesmiranda@gmail.com

Abstract. *Phishing websites are fake addresses that cheat the victims, passing by legitimate sites from banks or companies to obtain personal information without their consent. Looking to solve this problematic, several ways of defense were put into practice, among them the Machine Learning (ML). This article presents an study about ML utilization on malicious websites detection, reporting the methods used and concluding about their impact on precision and relevance.*

1. Introdução

Phishing é definido por criar falsos websites para se passar por endereços reais a fim de obter dados pessoais da vítima. Em [Banu and Banu 2013] está exemplificado alguns tipos de phishing, tais como clone phishing, onde hackers tentam imitar sites oficiais, alterando pequenos detalhes para dificultar a percepção; phone phishing, onde hackers se passam por falsas mensagens de bancos que chegam ao celular da vítima pedindo informações; e man-in-the-middle-attack, onde hackers interceptam mensagens verdadeiras que iam chegar no aparelho da vítima, e a modificam, inserindo malwares.

[Fette et al. 2007] afirma que primeira tentativa de conter ataques de phishing foi através de browse toolbars. Essa técnica atingia até 85% de eficiência, porém sua qualidade começou a cair devido à quantidade diminuída de informação contextual. Foi então que o foco mudou para a filtragem diretamente nos e-mails. Segundo [Abdelhamid et al. 2017], o Aprendizado de Máquina se provou bastante eficaz em diversos domínios, como diagnósticos médicos ou balança de mercado. Por isso, foi escolhido testar sua aplicabilidade na identificação de phishing. Neste mesmo artigo, também é realizada uma comparação entre classificadores, entre eles Support Vector Machines, Decision Trees e Neural Networks.

A seleção de atributos também se mostrou importante no combate ao phishing. Em [Fadheel et al. 2017] encontra-se um método de categorizar a sua relevância para detecção de sites maliciosos. Seus resultados mostraram uma lista dos melhores atributos selecionados, como HTTPS_token, having_IP_Address, Double_slash_redirecting, entre outros. Em [Li et al. 2019], além de atributos normais de URL, também foram selecionados novos atributos oriundos do HTML das páginas para reforçar a detecção.

Nesse artigo, apresentamos um estudo sobre cenários de avaliação para a identificação de páginas maliciosas. Buscamos avaliar classificadores a fim de identificar os mais relevantes, e realizamos um estudo mais profundo sobre a relevância de certos atributos nas bases de dados de phishing e sua capacidade de fornecer uma forte acurácia para a classificação.

2. Metodologia

Com o objetivo de estudar a eficiência dos modelos de classificação do AM para a resolução do problema do phishing, essa seção apresenta 5 cenários de avaliação, contendo cada uma um método alternativo que indique o impacto para a questão. As métricas foram submetidas em 6 datasets distintos, nomeados como [a], [b], [c], [d], [e] e [f], e disponibilizados[1].

Precisão e F1-score: Nessa aplicação, foram usados 4 datasets, sendo 3 deles com 30 atributos[a][b][c], e o outro com somente 9[d]. Os classificadores escolhidos para serem comparados quanto à eficácia foram Support Vector Machines (SVM), Decision Trees (DT) e Neural Network (NN), escolhidos devido à sua presença na literatura e fácil implementação. Para a obtenção da precisão e F1 Score foi usado o algoritmo *Classification Report*. O algoritmo testou os classificadores e avaliou sua capacidade de classificação, devolvendo ao final as respectivas respostas para cada um deles.

Grau de contribuição dos atributos: Nesta etapa, foram investigados o grau de contribuição de cada um dos atributos em cada um dos 4 datasets já utilizados, sendo utilizado o algoritmo Extremely Randomized Trees para obter os resultados. Ao final foi gerado um novo dataset com 12 atributos, resultado da junção de amostras das bases utilizadas mais os atributos melhores classificados pelo teste.

Comparação dos datasets: Tendo criado a base própria oriunda do método anterior, aplicamos novamente os algoritmos de classificação e o *Classification Report* para comparar seus resultados com os datasets já existentes. O resultado nos permitiu avaliar a escolha dos atributos em detrimento da pontuação obtida pela base.

Comparação de válidos e inválidos: Dois novos datasets foram considerados, um com 189 mil amostras de phishing válidos[e], e o segundo com 94 mil amostras de phishing inválidos[f], ambos extraídos do Phishtank[2]. Os dois possuíam os mesmos 12 atributos, sendo 10 com valores binários, então foi realizado um teste para comparar a quantidade de casos positivos em cada um desses 10 nos datasets. O resultado mostrou a importância de certos atributos devido à disparidade entre os válidos e inválidos.

Separação por tamanho da URL: A fim de simplificar o trabalho de classificação realizado pela máquina, foi analisado o atributo URL_Length, que possuía valores distintos entre si, para serem resumidos em 3 grupos: pequeno, médio e grande. Para isso, foi realizada uma média ponderada com a frequência de repetição dos grupos no teste.

3. Resultados Parciais

3.1. Classificador mais eficiente segundo Classification Report

Quando foram realizados os testes em todos os datasets, notamos que os que continham 30 atributos devolviam resultados iguais. Portanto, consideramo-os como um único resultado na resposta final. Como pode ser visto nas Tabelas 1 e 2, todos os classificadores apresentaram uma alta acurácia quando trabalhados em cima das bases fornecidas. Em todos os casos, Decision Trees se mostrou superior aos outros, estando sempre com um índice de 96% ou superior.

[1]Download dos 6 datasets: https://www.dropbox.com/s/cfrmtpfiajjp1e8/datasets.txt?dl=0

[2]https://www.phishtank.com

Classificador	30 atributos	9 atributos	Criado
SVM	0.95	0.88	0.94
DT	0.99	0.96	0.97
NN	0.94	0.80	0.92

Tabela 1. Precisão obtida pelo Classification Report

Classificador	30 atributos	9 atributos	Criado
SVM	0.95	0.87	0.94
DT	0.99	0.96	0.97
NN	0.94	0.83	0.92

Tabela 2. F1 Score obtido pelo Classification Report

3.2. Atributos mais relevantes segundo o grau de contribuição

Usando o Extremely Randomized trees para obter os atributos mais relevantes, observou-se que, assim como no método passado, os 3 datasets com 30 atributos devolveram resultados iguais. Os 9 atributos com maiores graus foram: URL_of_Anchor, SSLfinal_State, Preffix_Suffix, web_traffic, having_Sub_Domain, Links_in_tags, SFH, Request_URL e having_IP_Address. Já no dataset menor, os melhores classificados foram SFH e SSLfinal_State, tendo no total 6 atributos iguais aos das outras. Decidimos então criar a nova base mesclando esses 9 melhores das bases maiores com os 3 diferentes presentes na base menor, resultado em um dataset com 12 atributos.

3.3. Resultado da comparação entre os dataset

Como pode ser visto nas Tabelas 1 e 2, o teste por Classification Report revela a precisão e F1 Score médio das bases escolhidas. Como esperado, as bases com 30 atributos obtiveram os melhores resultados, com destaque para Decision Trees atingindo uma precisão de 99%. É visível, porém, que nossa base criada atingiu resultados similares segundo o teste, possuindo menos da metade dos atributos. A base menor, com 9 atributos, atingiu pontuações inferiores à nossa em todos os cenários possíveis.

3.4. Atributos mais relevantes entre phishing válidos e inválidos

Nesta etapa, realizamos um método manual de verificação de ocorrências. Alguns atributos, como having_IP_Address e SSLfinal_State continham poucas ou nenhuma ocorrência, o que mostra sua importância para a validação de phishing. Porém, sua ausência também é comum em phishings inválidos. Sendo assim, podemos observar os atributos que possuem uma alta discordância entre suas ocorrências, como having_Sub_Domain e Double_slash_redirecting, para reconhecê-los como determinantes para diferenciar entre phishings válidos e inválidos. O resultado pode ser visto na Tabela 3.

3.5. Categorização do tamanho das URLs

Neste método, foram feitas duas médias ponderadas, 1 para o dataset de phishings válidos e 1 para o de inválidos. Encontrou-se que o número médio de tamanho da URL para phishings válidos é 68,25. Já para phishings inválidos, o valor atingido foi 56,05. Decidiu-se então que URLs com tamanho acima de 60 fariam o atributo julgar como alto índice

Atributo	Válidos	Inválidos	Atributo	Válidos	Inválidos
having_IP_Address	0	2	SSLfinal_State	0	0
Shortining_Service	1621	576	PORT	977	958
having_At_Symbol	8327	1684	HTTPS_Token	0	0
Double_slash_redirecting	22215	3014	Redirect	0	0
having_Sub_Domain	107328	15567	Processed	189892(todos)	94785(todos)

Tabela 3. Quantidade de ocorrências

de phishing. Dessa forma, separou-se o tamanho em 3 categorias: pequeno (menores que 30), médio (entre 30 e 60) e grande (maiores que 60). Os valores pequenos foram substituídos por 1, os médios por 0 e os grandes por -1.

4. Conclusão e Trabalhos Futuros

Há diversos métodos de Aprendizado de Máquina para resolver o problema do phishing, porém, poucos conseguem ter um alto índice de precisão, seja por causa dos classificadores ou pelos atributos escolhidos. Sendo assim, esse artigo teve como objetivo estudar métodos para resolver a questão e avaliá-los com base nos seus resultados. Dentre os 3 classificadores escolhidos, Decision Trees foi o que mostrou melhor desempenho, sendo sempre superior em precisão.

Também foi mostrado que, apesar de possuírem uma grande relevância nas bases de phishing, atributos como having_IP_Address e SSLfinal_State também são frequentes em phishing inválidos, o que não favorece no impedimento dos falsos positivos. Assim, outros atributos como having_Sub_Domain provaram-se fundamentais pela grande diferença de ocorrência entre os sites maliciosos válidos e inválidos. Como trabalhos futuro, podemos citar o treinamento e teste com o dataset possuindo a URL_Length categorizada, e uma pesquisa sobre o uso de linguagem natural na identificação de phishings.

Referências

[Abdelhamid et al. 2017] Abdelhamid, N., Thabtah, F., and Abdel-jaber, H. (2017). Phishing detection: A recent intelligent machine learning comparison based on models content and features. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 72–77. IEEE.

[Banu and Banu 2013] Banu, M. N. and Banu, S. M. (2013). A comprehensive study of phishing attacks. *International Journal of Computer Science and Information Technologies*, 4(6):783–786.

[Fadheel et al. 2017] Fadheel, W., Abusharkh, M., and Abdel-Qader, I. (2017). On feature selection for the prediction of phishing websites. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing*, pages 871–876. IEEE.

[Fette et al. 2007] Fette, I., Sadeh, N., and Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web*, pages 649–656. ACM.

[Li et al. 2019] Li, Y., Yang, Z., Chen, X., Yuan, H., and Liu, W. (2019). A stacking model using url and html features for phishing webpage detection. *Future Generation Computer Systems*, 94:27–39.