Detecção de URLs Maliciosas

Mineração de Dados Aplicada

Julio Cesar da Silva Rodrigues[1]

[1]Universidade Federal de São João del-Rei
Curso de Ciência da Computação
julio.csr.271@aluno.ufsj.edu.br

13 de Junho de 2023

Roteiro

Roteiro

Status das Páginas

- URLs maliciosas possuem curto tempo de vida;
- Código HTTP retornado pode ajudar a destacar URLs deste tipo;
- Impacto observado na classificação foi positivo;
- Redução nos falsos negativos (*malware* classificado como seguro).

Holdout 80 | 20

XGBoost			
Class	Precision	Recall	F1-Score
Benign	0.93	0.98	0.95
Defacement	0.94	0.97	0.95
Phishing	0.96	0.89	0.93
Malware	0.94	0.82	0.88

Roteiro

Classe *benign*

- Correspondem à quase 70% do total da base;
- Classes *defacement* e *phishing* em relação à *benign* possuem:
  1. Macros F1 próximas;
  2. Cada uma apresenta menos que 25% da quantidade de instâncias.
- *Undersampling* aleatório;
- Remover grande parte destas instâncias, deixando alguma margem;
- Evitar perca de informação.

Classe *phishing*

- *PhishTank*[1] para suprir o déficit de instâncias de *phishing*;
- *Scraper* básico coletando URLs;
- Vantagem em relação à *oversampling* aleatório.

[1] Disponível em: `https://phishtank.org/phish_archive.php`

Classes *defacement* e *malware*

- Dificuldade em encontrar bases de dados com a quantidade de instâncias necessária;
- Utilização de *SMOTE* [Bowyer et al., 2011] em vez de *oversampling* aleatório;
- Gerar novas instâncias com base em dados existentes;

Base de Dados Final

- 800 mil instâncias, totalmente balanceada;
- 8 atributos e 1 classe (4 valores distintos);
- Composta por URLs de bases de dados do *Kaggle*[2] e *PhishTank*.

[2] Disponível em: `https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset`

Roteiro

Comparativo de Parciais

TP 2 - Parcial I		
Algoritmo	Média	Desvio Padrão
Regressão Logística	0,5780295368328738	0,0021293237223429956
XGBoost	0,856891093617907 9	0,0017065729226258411

TP 2 - Parcial II		
Algoritmo	Média	Desvio Padrão
Regressão Logística	0,67668635532121	0,0015789916679593162
XGBoost	0,9324923443237593	0,0008784997591179979

Teste t de dupla cauda

- Valores:
  1. $\alpha = 0,05$;
  2. $\mathrm{t} = 177,42420397526064$;
  3. $\mathrm{p} = 1,0791746370249748 \times 10^{-10}$.
- Hipótese nula rejeitada;
- Modelos estatisticamente distintos;
- XGBoost com nítida superioridade.

Roteiro

Faltou...

- Selecionar novos algoritmos para teste;
- Comparação com trabalhos relacionados;
- Busca por novos atributos?.

Referências


Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011).
SMOTE: synthetic minority over-sampling technique.
CoRR, abs/1106.1813.