Impacto do Aprendizado de Máquina Adversário contra Detectores de Anomalias em Séries Temporais

Felipe Dallmann Tomazeli, Gilberto Fernandes Junior, Bruno Bogaz Zarpelão

Departamento de Computação - Universidade Estadual de Londrina

Universidade Estadual de Londrina

ARAUCÁRIA
CIÊNCIA, TECNOLOGIA E INOVAÇÃO PARA O PARANÁ

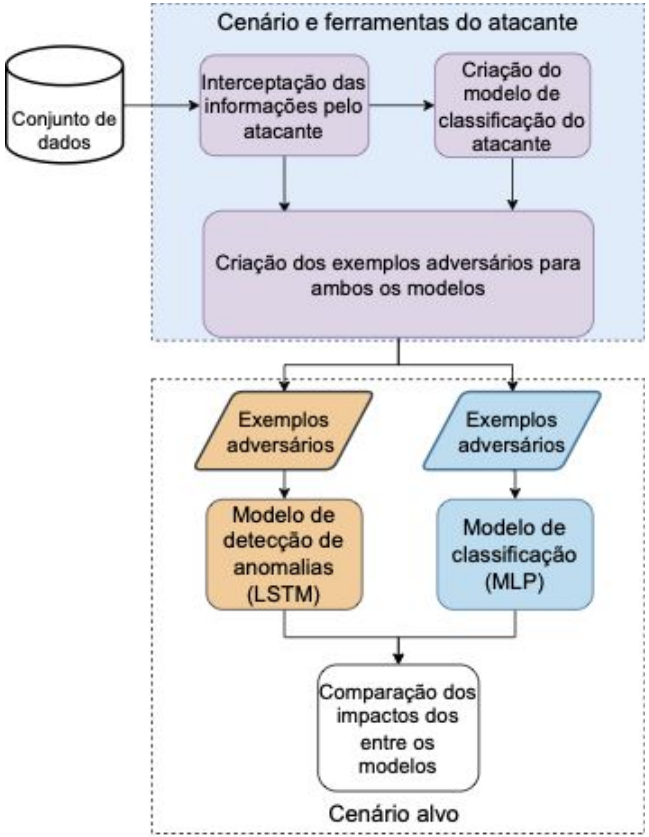
Motivação

- Monitoramento de séries temporais é utilizado em aplicações críticas.
- Algoritmos de aprendizado de máquina podem ser vulneráveis a exemplos adversários:
  - Temos ampla literatura sobre ataques contra classificadores baseados em redes neurais.

Objetivo

- Qual seria o impacto de exemplos adversários, criados com métodos voltados a classificadores, contra um detector de anomalias em séries temporais?

Visão geral do experimento



Conjunto de dados

- Público e disponível na Web.
- Informações coletadas de um sistema de circulação de água.
- 8 dimensões (*features*).
- 751 observações, sendo 557 normais e 194 anômalas.

Métricas de avaliação

- Precisão: de todas as anomalias detectadas, quantas realmente ocorreram.
- Revocação: de todas as anomalias ocorridas, quantas foram detectadas.
- F1-score: média harmônica entre precisão e revocação.
- Mathews Correlation Coefficient: calcula a correlação entre as predições e as respostas reais.

Modelo alvo: detecção de anomalias

- Dados divididos em treinamento, definição de limiar e teste.
- LSTM treinada para receber série de 5 observações e predizer a sexta.
- Limiar é o erro quadrático médio das predições somada ao desvio padrão dos erros quadráticos multiplicado por um fator *z*.
- Inferência: distância Euclidiana entre predição e observação real é comparada ao limiar.

Modelo alvo: classificador

- Dados são divididos em treinamento e teste (60% - 40%).
- Deve-se garantir a presença de amostras anômalas e normais nos dos subconjuntos.
- O classificador é uma rede neural MLP com 8 entradas (uma para cada dimensão do conjunto de dados) e uma saída.

Modelo adversário: classificador

- Criação de exemplos adversários baseados no método FGSM dependem de um classificador treinado sobre os dados usados pelo alvo.
- O ataque pressupõe, portanto, que o atacante tem acesso aos dados de treinamento utilizados pelo alvo.
- O modelo adversário é baseado em um classificador MLP idêntico ao utilizado pelo alvo.

Exemplos adversários

- Entradas manipuladas com o menor grau de perturbação possível para maximizar o erro do modelo alvo.
- Ataque naive: soma-se, a cada feature, o desvio padrão daquela feature multiplicado por um fator *k* (determina a severidade do ataque).

Exemplos adversários

- Ataque FGSM: perturbação adicionada à amostra depende do gradiente do modelo classificador treinado. Há também um fator epsilon que define a severidade do ataque.
- Nos dois ataques, perturbações são adicionadas ou subtraídas a depender da intenção:
  - Induzir falsos positivos ou falsos negativos.

Resultados - Gerar falsos positivos no LSTM

	naive		FGSM	
severidade	**precisão**	**TFP**	**precisão**	**TFP**
sem ataque	0,96	0,06	0,96	0,06
máxima	0,93	0,12	0,74	0,56

Resultados - Gerar falsos negativos no LSTM

	naive		FGSM	
severidade	**revocação**	**TFN**	**revocação**	**TFN**
sem ataque	0,93	0,06	0,93	0,06
média	0,91	0,08	0,86	0,13
máxima	0,89	0,10	0,93	0,06

Resultados - Gerar falsos positivos no MLP

	naive		FGSM	
severidade	**precisão**	**TFP**	**precisão**	**TFP**
sem ataque	0,98	0,00	0,98	0,00
máxima	0,88	0,04	0,25	0,94

Resultados - Gerar falsos negativos no MLP

	naive		FGSM	
severidade	**revocação**	**TFN**	**revocação**	**TFN**
sem ataque	0,94	0,05	0,94	0,05
máxima	0,37	0,62	0,05	0,94

Considerações finais

- O modelo de detecção de anomalias baseado em LSTM foi mais robusto contra os ataques.
- Ataques do tipo FGSM demonstraram maior poder de diminuir capacidade preditiva dos alvos.

Trabalhos futuros

- Desenvolver técnicas de defesa voltadas a séries temporais.

Obrigado!

- Bruno B. Zarpelão
- brunozarpelao@uel.br

Patrocinadores do SBSeg 2024!

nic.br

cgi.br

Google

Tempest

CAPES

SiDi

FAPESP

zscaler™

BugHunt

CNPq

C.E.S.A.R

FACULDADE IBPTECH