Survey em IA e Segurança

Davi Iury, Esther Martins, Lucas Pinheiro, Rafael Porto, and Théo Araújo

Universidade Federal do Ceará

Abstract. Nós resume as coisas aqui.

Keywords: IA · Segurança · Machine Learning.

1 Introdução

IA é muito popular. Porém, precisamos de muitos dados para treinar modelos. Como podemos arranjar esses dados? Mais especificamente, como podemos arrumar esses dados de forma que não infrijamos leis de privacidade de dados? Como privacidade de dados vem se tornando um conceito cada vez mais em voga, Novas formas de treinar modelos e obter dados vem surgindo. Nesta survey, falaremos de:
Federated learning (analisar os dados locamente e mandar os resultados de volta de forma criptografada)
Differential privacy (é uma técnica que visa proteger a privacidade dos usuários por meio da adição de ruído nos dados sendo analisados)
Machine Unlearning (esquecer dados de usuários que foram usados para treinar modelos de forma que isso não prejudique o aprendizado do algoritmo).

1.1 Contextualização

Deixei aqui para ser o template inicial. Quando forem escrever, É legal dar enter a cada oração Para que fique dividido direito e fique fácil de ler.

2 Caracterização Ferramental

2.1 Machine unlearning

woow

2.2 Differential privacy

woow

2.3 Federated learning

Contexto. É comum que algoritmos clássicos de aprendizado de máquina mantenham centralizados os dados a serem usados para treinamento. Isso se deve ao fato de que, frequentemente, os dados encontram-se dispersos em ¨ilhas de dados¨[1], então, é necessário que seja feito um trabalho de captação e agrupamento em um servidor para que o uso em treinamento seja possível. No entanto, caso não seja feita de forma adequada, essa centralização facilita o vazamento de dados sensíveis. Em vista dessa situação, diversas regulações vem sendo impostas com relação à captação e ao uso de dados para treinamento de modelos, tornando o uso de técnicas de aprendizado de máquina centralizado de difícil implementação prática. Nesse contexto, a aplicação do federated learning possibilita com que o treinamento seja feito de forma local, não-centralizada, de forma que os dados de um usuário específico mantém-se somente no seu dispositivo local.
Definindo. Em sua essência, o federated learning é uma técnica de aprendizado distribuído, ou seja, ao invés de consolidarmos dados de usuário em um servidor central para treinar um modelo, haverão focos locais de treinamento. Assim, evitando a captação de dados sensíveis, o servidor envia um modelo de treinamento para cada dispositivo individual, mantendo a fase de treinamento como uma etapa local. Cada dispositivo somente retornará ao servidor central seu modelo treinado localmente e atualizará o seu modelo interno conforme as atualizações no modelo global. Federated learning, dessa forma, garante que dados locais não possam ser vazados e, a fim de não violar leis gerais de proteção de dados, a troca de parâmetros entre clientes locais e o servidor para a geração de um modelo global é feita através de mecanismos criptografados.[2]
Tipos.
1. Com relação à partição dos dados.
a. Sistemas que fornecem os dados tem usuários muitos diferentes, mas features parecidas b. Sistemas tem features muito diversas, mas usuários muito parecidos c. Sistemas fornecem dados muito incompatíveis (features e usuários diferentes) ou insuficientes.
2. Com relação à mecanismos de privacidade

3. Com relação ao modelo de ML aplicado

4. Com relação ao método de solucionar heterogênidade.

Aplicações práticas.

3 Geração de Dados Sintéticos e Segurança

A escassez de dados de alta qualidade, combinada com regulamentações rigorosas de privacidade e o alto custo de anotação manual, impulsionou a adoção

[1] to-do referencia.
[2] botar uma figura aqui

da geração de dados sintéticos como uma alternativa viável aos *datasets* reais [?]. Dados sintéticos são definidos como informações artificialmente geradas por algoritmos ou simulações que mimetizam as propriedades estatísticas e comportamentais dos dados reais, sem conter informações diretamente identificáveis. No contexto de sistemas de Inteligência Artificial seguros e privados, a geração de dados sintéticos não atua apenas como uma ferramenta de aumento de dados (*data augmentation*), mas como um mecanismo fundamental de preservação de privacidade e robustez [?].

3.1 Técnicas de Geração Baseadas em IA Generativa

A literatura recente categoriza as abordagens de geração de dados sintéticos principalmente em três arquiteturas de aprendizado profundo:

- **Redes Adversárias Generativas (GANs):** Consistem em dois modelos neurais, um gerador e um discriminador, que competem entre si. O gerador cria dados falsos e o discriminador tenta distinguir entre dados reais e falsos. GANs são amplamente utilizadas para síntese de imagens e dados tabulares, com variantes aplicadas especificamente para preservar a estrutura estatística de dados sensíveis [?].
- **Autoencoders Variacionais (VAEs):** Os VAEs aprendem a comprimir os dados de entrada em um espaço latente probabilístico e, em seguida, reconstróem os dados a partir desse espaço, permitindo a amostragem de novos pontos de dados que seguem a distribuição original [?].
- **Grandes Modelos de Linguagem (LLMs):** Avanços recentes, como apontado por Nadăş et al. [?], demonstram que LLMs podem atuar como geradores de dados universais para texto e código. Através de técnicas de *prompting* e refinamento iterativo, LLMs podem gerar dados rotulados e código sintético verificável via execução, auxiliando no treinamento de modelos de segurança de software.

3.2 Sinergia com Mecanismos de Privacidade e Segurança

A geração de dados sintéticos desempenha um papel estratégico quando integrada às técnicas de *Differential Privacy*, *Federated Learning* e *Machine Unlearning*.

Privacidade Diferencial (DP). A aplicação direta de modelos gerativos em dados sensíveis pode resultar em riscos de privacidade, como ataques de inferência de pertinência (*membership inference attacks*). Para mitigar isso, a geração de dados sintéticos é frequentemente acoplada à Privacidade Diferencial [?]. O objetivo é gerar um *dataset* sintético que preserve as propriedades estatísticas globais, mas garanta que a saída do modelo não dependa excessivamente de nenhum registro individual. Técnicas como DP-GAN integram ruído calibrado garantindo garantias formais de privacidade (ϵ-differential privacy), permitindo que dados sintéticos sejam compartilhados publicamente com risco minimizado.

Aprendizado Federado (FL). No Aprendizado Federado, a heterogeneidade dos dados (dados *non-IID*) é um desafio crítico. A geração de dados sintéticos oferece soluções em duas frentes [**?**]:

1. **Aumento de Dados Local:** Clientes com poucos dados podem usar modelos generativos para sintetizar amostras locais, melhorando a robustez do treinamento.
2. **Privacidade Generativa Federada:** O treinamento de GANs em ambiente federado, onde apenas os parâmetros dos geradores são compartilhados, permite que o modelo global aprenda a distribuição de dados sem acesso direto aos registros brutos.

Machine Unlearning. O *Machine Unlearning* visa remover a influência de dados específicos de um modelo treinado. A geração de dados sintéticos atua como facilitador neste processo:

- **Prevenção de Memorização:** O uso de dados sintéticos treinados com DP desde o início reduz a necessidade de *unlearning* frequente, pois os dados não correspondem a indivíduos reais.
- **Substituição de Dados:** Em cenários onde dados reais devem ser excluídos, dados sintéticos podem ser gerados para preencher a lacuna estatística deixada pela remoção, mantendo a utilidade do modelo sem violar a privacidade [**?**].

4 Desafios