Defendendo Modelos de Aprendizado de Máquina: Estratégias de Detecção de Anomalias Contra Ataques de Envenenamento

Livia Hipolito[1], **Amanda Lopes**[1], **Manoel Clipes**[1], **Lucas Bastos**[1]

[1]Federal University of Pará (UFPA) - Belém, PA, Brazil

{livia.pires, amanda.lopes,manoel.clipes}@itec.ufpa.br,

lucas.bastos@itec.ufpa.br

Resumo. *Este trabalho apresenta o framework de defesa GAIA, desenvolvido para detectar e mitigar ataques de envenenamento por parâmetros aleatórios em Aprendizado Federado (FL). Tais ataques, os quais envolvem a submissão de tensores com valores aleatórios por agentes maliciosos, comprometem a integridade e a acurácia dos modelos globais. A metodologia utiliza as normas vetoriais L1 e L2 para extrair características das atualizações de modelo. As atualizações de clientes benignos seguem padrões estatisticos consistentes, enquanto as de atacantes exibem distribuições divergentes. A análise dessas assinaturas permite ao servidor de agregação identificar atualizações maliciosas e garantir a estabilidade e precisão do modelo federado. Por meio de demonstração experimental, a eficácia da estrutura proposta foi validada ao atingir 40% de acurácia, o que representa um aumento de mais de 30% em relação ao modelo sem defesa.*

Abstract. *This work presents the GAIA defense framework to detect and mitigate poisoning attacks using random parameters in Federated Learning (FL). These attacks, involving the submission of tensors with random values by malicious agents, compromise the integrity and accuracy of global models. The methodology uses L1 and L2 vector norms to extract features from the model updates. Updates from benign clients follow consistent statistical patterns, while those from attackers exhibit divergent distributions. Analyzing these signatures allows the aggregation server to identify malicious updates and ensure the stability and accuracy of the federated model. Through experimental demonstration, the effectiveness of the proposed framework was validated by achieving 40% accuracy, representing an improvement of over 30% compared to the model without defense.*

1. Introdução

Ao longo dos anos, com o fortalecimento da LGPD (Lei Geral de Proteção de Dados) e o contínuo avanço no desenvolvimento de redes neurais, tem-se observado um aumento de demanda sem precedentes. No entanto, a necessidade de treinar modelos sem o compartilhamento de dados tornou-se um grande desafio para empresas que desenvolvem redes neurais, uma vez que seus clientes relutam em fornecer dados sensíveis [Korkmaz et al. 2022]. O Aprendizado Federado (do inglês, Federated Learning – FL) surge como uma

solução, sendo uma abordagem de treinamento de modelos de aprendizado de máquina que opera em um paradigma descentralizado. Diferentemente das metodologias centralizadas tradicionais, as quais exigem a agregação de dados em um servidor único para processamento, o FL permite que o treinamento ocorra diretamente nos dispositivos locais (ou nós), onde os dados são originalmente gerados e armazenados [AbdulRahman et al. 2020].

Ao modificar o método tradicional de treinamento de máquinas, surgem novos desafios para garantir a generalização da tarefa treinada pela rede [de Souza et al. 2024]. A ameaça de dados falsos ou manipulados durante o treinamento local é uma preocupação crítica que afeta diretamente a confiabilidade e a integridade do modelo final em sistemas de aprendizado de máquina [Blanchard et al. 2017]. Essa vulnerabilidade é particularmente acentuada em contextos de FL, onde a descentralização do treinamento pode introduzir desafios adicionais. A falta de monitoramento adequado na seleção dos clientes participantes agrava a situação, permitindo que atores mal-intencionados injetem dados corrompidos ou tendenciosos sem detecção.

Uma abordagem em menor escala pode ser uma solução eficaz para reduzir a quantidade de dados a serem comparados e para a exclusão de clientes maliciosos [de Souza et al. 2023]. Nesse cenário, algumas abordagens utilizam a clusterização como método principal para classificar clientes como benignos ou maliciosos. No entanto, esses métodos geralmente enfrentam limitações de escalabilidade. Em contraste, o uso de normas vetoriais, como as normas L1 e L2, que descrevem características específicas dos clientes, oferece uma alternativa eficiente, contribuindo para a atualização do modelo de forma mais escalável e robusta.

Este trabalho apresenta o algoritmo de defesa GAIA, projetado para detectar e mitigar ataques de envenenamento com parâmetros aleatórios em FL. Nesses ataques, agentes maliciosos enviam tensores aleatórios para comprometer a acurácia e a integridade do modelo global. GAIA utiliza as normas vetoriais L1 e L2 para gerar uma ”assinatura”que resume a magnitude das atualizações. Atualizações benignas seguem um padrão estatístico estável, enquanto ataques exibem desvios detectáveis. Com base nessas assinaturas, o servidor identifica e descarta contribuições maliciosas, preservando a convergência e a qualidade do modelo.

O restante deste trabalho está organizado da seguinte forma: a Seção 2 apresenta uma visão geral dos trabalhos relacionados à detecção de clientes maliciosos e ao FL. A Seção 3 descreve nossa metodologia para a detecção de clientes maliciosos. A Seção 4 explora os resultados alcançados neste trabalho. Finalmente, a Seção 5 apresenta as conclusões deste artigo.

2. Trabalhos Relacionados

Morais *et al.* no campo da segurança em FL, a análise de normas de parâmetros tem se mostrado uma estratégia promissora. Utilizando um conjunto tridimensional de características, calculando a norma L3 para cada atualização de cliente. Subsequentemente, aplica algoritmos de clusterização sobre esses vetores de características com o objetivo de segregar os agentes em grupos distintos de ”benignos”e ”maliciosos”. Em contraste, nossa pesquisa adota uma perspectiva diferente e mais fundamental [Morais et al. 2024].

Assumção e Villas apresentam uma abordagem de FL que otimiza o treinamento

em cenários com dados não-IID (não independentes e identicamente distribuídos) por meio da seleção estratégica de clientes. Além disso, o trabalho propõe um método de avaliação de treinamento compatível com criptografia homomórfica, garantindo um monitoramento seguro e privado. Essa funcionalidade é crucial para preservar o progresso do treinamento em caso de um ataque de envenenamento, garantindo que o conhecimento adquirido não seja perdido [Assumpção and Villas 2024].

Em vez de combinar múltiplas normas para clusterização, nosso trabalho foca em uma avaliação empírica e comparativa da eficácia individual das normas L1 e L2 como detectores diretos de envenenamento. O objetivo central é usar estas duas propostas de regularização como termos únicos para classificar distorções no processo de treino,, desta forma detectando clientes indesejáveis ou maliciosos da agregação.

3. GAIA

Esta seção apresenta o algoritmo GAIA, que identifica ataques realizados por clientes maliciosos em ambientes de FL, nos quais esses clientes enviam atualizações com valores que não contribuem para o aprendizado da rede. O algoritmo utiliza as normas vetoriais como base para classificar de forma eficiente os clientes que não agregam conhecimento útil à rede, permitindo uma análise ordenada. Esta seção descreve o modelo do sistema e os detalhes operacionais do GAIA.

3.1. Visão geral do cenário

Consideramos um cenário com n dispositivos $\mathcal{U} = \{u_1, \ldots, u_n\}$, onde, a cada rodada de FL, é selecionado um subconjunto $\mathcal{C} \subseteq \mathcal{U}$ para treinar o modelo global M_g com seus dados locais D_i. O mecanismo de seleção escolhe os clientes cujos dados têm maior impacto no treinamento do modelo, ajustando seus respectivos modelos locais M_i com base nesses dados. A agregação dos modelos locais em um modelo global é feita utilizando a abordagem FedAvg, que calcula a média ponderada dos parâmetros θ_{global} conforme a Eq. 1, com os pesos w_i determinados pelo tamanho dos conjuntos de dados $|D_i|$ de cada cliente.

$$\theta_{global} = \frac{1}{n} \sum_{i=1}^{n} w_i \theta_i \tag{1}$$

Aqui, w_i representa o peso atribuído a cada cliente com base no tamanho de seu conjunto de dados D_i, enquanto θ_i são os parâmetros locais treinados por cada cliente, calculados conforme a Eq. 2. Dessa forma, o FedAvg ajusta o modelo global levando em conta a contribuição proporcional de cada cliente, com base no tamanho de seus dados.

$$w_i = \frac{|D_i|}{\sum_{i=1}^{n} |D_i|} \tag{2}$$

3.2. Funcionamento da prevenção de ataques

Este trabalho foca em identificar diferenças entre clientes com base em seus modelos. Para detectar clientes maliciosos, é essencial reconhecer mudanças abruptas ou padrões fora do comum. Em termos matemáticos, isso se assemelha a um sistema comparativo, onde interseções e distâncias entre conjuntos destacam comportamentos atípicos.

Nesta proposta, a abordagem inspirada no trabalho de [Morais et al. 2024], utilizando as normas L1 e L2, mostrou-se muito útil para comparar a distorção dos dados de um modelo quando comparados com um cluster maior, definindo, assim, critérios para a exclusão de dados discrepantes. A norma L1 mede a "distância" total ao longo dos eixos coordenados, sendo bastante eficaz para problemas de otimização e aprendizado de máquinas. A regularização L1 é dada pela soma dos valores absolutos de seus componentes, conforme mostrado na equação 3.

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \tag{3}$$

Agora, a função L2 é dada pela raiz quadrada da soma dos quadrados dos seus componentes, como mostrado na equação 4. A norma L2 calcula a distância euclidiana de um vetor ao ponto de origem no espaço, ou seja, ela representa a raiz quadrada da soma dos quadrados de cada componente do vetor.

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{4}$$

O cálculo das raízes da matriz e das distâncias entre seus parâmetros ajuda a identificar distorções e vieses na rede. Esses indicadores facilitam a escolha dos clientes e revelam comportamentos anômalos, como no caso dos ataques com valores randômicos simulados.

O *GAIA* utiliza o algoritmo *K-means* para agrupar atualizações de modelo em *clusters* com comportamentos semelhantes. A expectativa é que clientes maliciosos, por apresentarem atualizações divergentes, formem um *cluster* separado. O *K-means* agrupa os dados $X = \{x_1, x_2, \ldots, x_n\}$ em k *clusters*, minimizando a soma das distâncias quadráticas aos centros, conforme a equação 5. Usando os vetores calculados com L1 e L2, são formados os *clusters* que indicam o aprendizado de cada modelo. O *cluster* com menos clientes deve conter os destoantes, que serão removidos da agregação.

$$J = \sum_{k=1}^{k} \sum_{x_i \in C_k} \|x_i - c_k\|^2 \tag{5}$$

4. Avaliação

Nesta seção, apresentamos os resultados das simulações realizadas com 100 rodadas, 50 clientes e uma taxa fixa de 10% de clientes maliciosos, ajustável conforme necessário usando o dataset de imagens CIFAR-10 contendo 10 classes diferentes de 32x32 pixels, em cores RGB. Foram utilizados métodos de classificação baseados nas normas L1 e L2, além de um cenário sem defesa para validar a eficácia dos ataques e servir como referência ideal de FL sem comprometimento. As quatro simulações representam os principais métodos utilizados na validação da proposta, permitindo comparar a acurácia e a perda em cenários de FL sob ataque.

A Figura 1 mostra os resultados de acurácia. No cenário ideal, sem ataques, a agregação dos modelos foi consistente, alcançando valores acima de 60% de acurácia em 40 rodadas, enquanto o mesmo modelo, sofrendo os ataques, não consegue evoluir mais de 10%. Os modelos de defesa L1 e L2 apresentaram cerca de 40% de acurácia, superando o modelo sem defesa, que teve um aumento de mais de 30% em relação ao modelo com ataques. Enquanto o modelo sem defesa não evoluiu, os modelos com defesa mostraram evolução constante, evidenciando a eficácia da defesa.

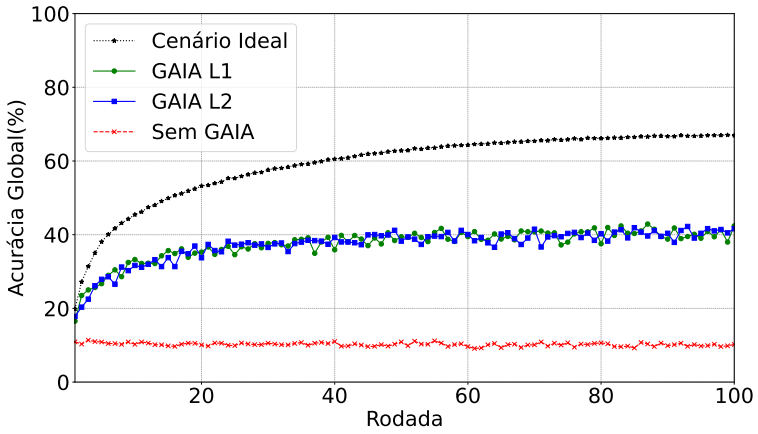

Figura 1. Medição de Acurácia

A Figura 2 mostra que no cenário sem defesa, os resultados indicam o desaprendizado que pode ocorrer quando o servidor de agregação recebe um ataque, o que gera vários saltos, com valores de perda superiores a 10 em diversos casos. Em outros cenários, como o de defesa, observamos uma estagnação devido à quantidade de clientes agregados. Ou seja, a quantidade de clientes não aprende tanto, mas não há um método de desaprendizado, já que os clientes aprendem e mantêm valores de perda abaixo de 2. Além disso, o cenário ideal, onde não há ataque, apresenta perda inferiores a 1, indicando um aprendizado quase completo em relação ao *dataset* utilizado.

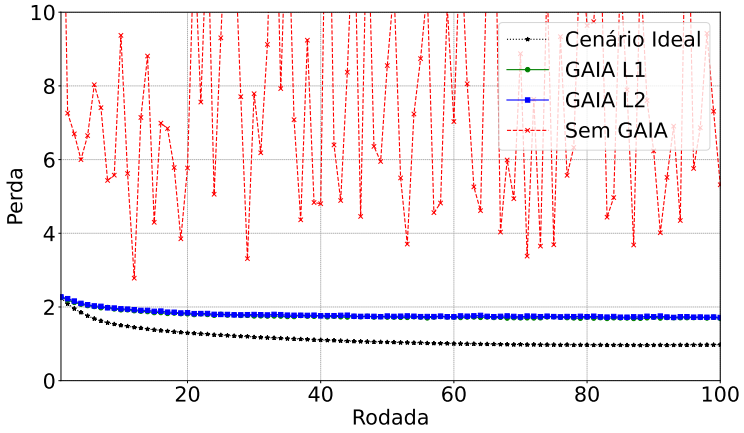

Figura 2. Medição de Perda

5. Conclusão

Este trabalho propôs uma abordagem para mitigar ataques de envenenamento de modelos em ambientes de FL , utilizando normas vetoriais L1 e L2 como critérios de defesa. O algoritmo GAIA demonstrou elevada eficácia na identificação e mitigação de atualizações maliciosas, assegurando a integridade e a precisão do modelo global. As simulações evidenciaram que, mesmo sob ataques, os mecanismos de defesa baseados nas normas L1 e L2 superaram significativamente o desempenho do modelo sem proteção, alcançando até 30% de melhoria na acurácia em comparação ao cenário atacado sem defesa. No cenário ideal, sem ataques, os resultados também validaram a efetividade da abordagem, embora tenham revelado gargalos que ainda precisam ser superados como a necessidade de estratégias complementares, por exemplo, uma seleção mais criteriosa de clientes para o treinamento.

Como trabalho futuro, buscamos aprimorar o algoritmo GAIA com a implementação de métodos adaptativos de agregação local e explorar sua aplicação em cenários FL mais dinâmicos, com fluxos de dados variáveis para cada cliente. A integração desses métodos visa aumentar ainda mais a robustez e a eficiência do aprendizado federado em ambientes sujeitos a ataques.

Agradecimentos

Esta pesquisa foi financiada pela Fundação Amazônia para Estudos e Apoio à Pesquisa (FAPESPA).

Referências

AbdulRahman, S., Tout, H., Ould-Slimane, H., Mourad, A., Talhi, C., and Guizani, M. (2020). A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. *IEEE Internet of Things Journal*, 8(7):5476–5497.

Assumpção, N. R. and Villas, L. A. (2024). Rápido, privado e protegido: Uma abordagem para aprendizado federado eficiente em ambiente hostil. In *Workshop de Computação Urbana (CoUrb)*, pages 15–28. SBC.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30.

de Souza, A. M., Bittencourt, L. F., Cerqueira, E., Loureiro, A. A., and Villas, L. A. (2023). Dispositivos, eu escolho vocês: Seleçao de clientes adaptativa para comunicaçao eficiente em aprendizado federado. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, pages 1–14. SBC.

de Souza, A. M., Maciel, F., da Costa, J. B., Bittencourt, L. F., Cerqueira, E., Loureiro, A. A., and Villas, L. A. (2024). Adaptive client selection with personalization for communication efficient federated learning. *Ad Hoc Networks*, 157:103462.

Korkmaz, A., Alhonainy, A., and Rao, P. (2022). An evaluation of federated learning techniques for secure and privacy-preserving machine learning on medical datasets. In *2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE.

Morais, M. G., da Costa, J. B., Gonzalez, L. F., de Souza, A. M., and Villas, L. A. (2024). Mecanismo para mitigar ataques de envenenamento de modelo no aprendizado federado. In *Workshop de Computação Urbana (CoUrb)*, pages 224–237. SBC.