SCOPE-FL:Seleção de Clientes por Ordem de Entropia

Isaque O. Silva[1], **Carlos Vitelli**[1], **Iago Medeiros**[1]

[1]Federal University of Pará (UFPA)

isaque.silva@tucurui.ufpa.br, carlos.vitelli@itec.ufpa.br,

iagomedeiros@ufpa.br

Resumo. *A crescente utilização de dispositivos conectados exige novos métodos para lidar com a quantidade e privacidade dos dados compartilhados. Federated Learning (FL) surge como uma solução, permitindo o treinamento de modelos sem compartilhar dados diretamente, preservando a privacidade dos clientes. No entanto, nem todos os clientes são igualmente úteis para o aprimoramento de modelos globais, tornando necessária uma seleção eficiente de clientes. O SCOPE-FL propõe um mecanismo dinâmico de seleção de clientes, atribuindo pesos à entropia dos dados e ao tamanho do dataset, para garantir uma contribuição mais eficiente para o modelo global. Isso é feito calculando uma pontuação de relevância para cada cliente, com base nesses fatores, e ajustando os pesos atribuídos a cada cliente. O SCOPE-FL usa o método FedAvg para agregar modelos locais, priorizando clientes com dados mais relevantes. Apos simulações utilizando o dataset MNIST, o SCOPE-FL superou métodos tradicionais, mostrando uma taxa de acurácia superior a 60% após 12 rodadas, alcançando até 80% em 22 rodadas.*

Abstract. *The increasing use of connected devices requires new methods to handle the quantity and privacy of shared data. Federated Learning (FL) emerges as a solution, enabling model training without directly sharing data, preserving the clients' privacy. However, not all clients are equally useful for improving global models, making efficient client selection necessary. SCOPE-FL proposes a dynamic client selection mechanism, assigning weights to data entropy and dataset size to ensure a more efficient contribution to the global model. This is done by calculating a relevance score for each client based on these factors and adjusting the weights assigned to each client. SCOPE-FL uses the FedAvg method to aggregate local models, prioritizing clients with more relevant data. Tested with MNIST, SCOPE-FL outperformed traditional methods, showing an accuracy rate of over 60% after 12 rounds, reaching up to 80% after 22 rounds.*

1. Introdução

O crescente avanço de novas tecnologias, como carros autônomos, dispositivos móveis, entre outros, transformou o cotidiano das pessoas. A expectativa atual é de que, até 2030, existam mais de 500 bilhões de dispositivos conectados no planeta. Isso representa uma quantidade colossal de dispositivos alocados. Devido à conectividade de dispositivos cada vez menores, a rede, a distribuição e a limitação dos dados compartilhados por esses dispositivos tornam-se ainda mais arriscadas. Esse fenômeno facilita a integração de ataques aos dados e invasões que contribuem para a perda de dados sensíveis dos usuários. Devido à ineficiência do aprendizado de máquina sem a disponibilidade de dados, são necessários novos métodos para escalonar predições e classificações de dispositivos na

borda. No entanto, as aplicações de Machine Learning (ML) exigem um compartilhamento significativo de dados, o que traz importantes questões de comunicação e privacidade [AbdulRahman et al. 2020].

Para solucionar o problema de risco relacionado ao compartilhamento e à quantidade de dados enviados no processo de aprendizado de máquina, foi desenvolvida uma metodologia que leva em conta apenas os pesos do modelo, conhecida como *Federated Learning* (FL), também chamado de Aprendizado Federado [McMahan et al. 2017]. Essa proposta visa eliminar o armazenamento indevido de dados para o treinamento de máquina, enquanto mantém a privacidade dos clientes. Essa abordagem facilita a generalização dos dados, colocando o desenvolvimento do aprendizado nos dispositivos de borda, locais onde atuam, e diminuindo a quantidade necessária de dados que o cliente envia para um servidor.

No entanto, no cenário de FL, nem todos os clientes são necessariamente ideais para o desenvolvimento e aprimoramento das aplicações. Em alguns casos, pode ser necessário trabalhar com um número reduzido de clientes devido à relevância dos dados que eles fornecem. O papel central do FL na melhoria da acurácia durante o treinamento de aprendizado de máquina está no foco de generalização do modelo global, de forma que qualquer novo participante possa utilizar o modelo global atualizado [Pires et al. 2020]. Portanto, ao escolher os clientes, variando com base na quantidade do seu *dataset* ou no desempenho do seu treinamento em relação à perda do modelo global, temos argumentos que são interdependentes para aprimorar a arquitetura. Utilizar menos clientes pode excluir a necessidade de mais iterações, permitindo que o modelo se estabilize mais rapidamente.

Neste artigo, buscamos trabalhar com a seleção de clientes de forma mais genérica, a fim de evitar exclusões desnecessárias e garantir uma relativa generalização do modelo. Ou seja, o SCOPE-FL auxiliará o servidor a escolher os clientes sem necessariamente excluir sua participação, mas atribuindo pesos para que sempre haja a possibilidade de o cliente ser selecionado. Essa abordagem contribui tanto para a melhoria da acurácia quanto para a generalização, pois, ao atribuir pesos à Entropia dos dados do cliente e relacionar um peso secundário ao tamanho do seu *dataset*, garantimos uma especialização ao treinar os clientes de forma genérica e asseguramos uma estrutura mais palpável desse cenário.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta uma visão geral dos trabalhos sobre seleção de clientes. A Seção 3 descreve nossa metodologia para a conexão servidor-cliente e a abordagem SCOPE-FL. A Seção 4 explora o modelo de simulação e os resultados relacionados ao nosso método. Finalmente, a Seção 5 conclui o artigo.

2. Trabalhos Relacionados

Souza *et al.* usa uma abordagem de diferenciação entre os clientes de acordo com seus desempenhos estruturais. Dessa forma, a utilização de recursos pode ser reduzida exponencialmente, diminuindo tanto a quantidade de clientes que treinam por rodada quanto a variabilidade da generalização, ao variar os clientes selecionados. Esta abordagem proporciona uma maior integração entre os clientes, tornando-os menos repetidos e utilizando um cenário mais dinâmico e menos estocástico [de Souza et al. 2023].

Deng *et al.* utiliza uma abordagem mais relativista e estática em relação à seleção

de clientes. Enquanto o aprendizado funciona normalmente, o servidor verifica quais clientes apresentam maior variabilidade na qualidade. Esse fator pode variar tanto em função da quantidade de rótulos quanto do tamanho total dos dados, funcionando como um método de ranqueamento para a seleção de clientes [Deng et al. 2021].

Esses trabalhos têm como principal foco a seleção de clientes, e ambos se preparam para selecionar com base na escolha estrutural dos clientes, a fim de realizar a decisão de prioridade a clientes a serem escolhidos para realizar a próxima etapa de treino. Dessa forma, este trabalho foca em tornar o cenário dinâmico, utilizando os dois métodos para selecionar os clientes de maneira não estocástica.

3. SCOPE-FL

Esta seção introduz o algoritmo do SCOPE-FL, o qual melhora a seleção de clientes em relação ao estado da arte no processo de FL ao considerar os valores de Entropia que aquele modelo gera e a quantidade de dados de cada cliente. O algoritmo utiliza como base os clientes com maior quantidade de dados para ranqueá-los e, em seguida, verifica os valores de peso, mensurando que, quanto maior o peso, mais preferível é a escolha desse cliente a cada interação a partir da entropia de Shannon de seus dados [Orlandi et al. 2023]. Esta seção descreve o modelo do sistema e os detalhes operacionais do SCOPE-FL.

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, um subconjunto $\mathcal{C} \subseteq \mathcal{U}$ é selecionado para treinar o modelo global M_g com seus dados locais D_i. O mecanismo de seleção escolhe os clientes cujos dados mais contribuem para o treinamento, ajustando seus modelos locais M_i com base em seus conjuntos de dados. A abordagem FedAvg é utilizada para agregar os modelos locais em um modelo global, calculando a média ponderada dos parâmetros θ_{global} a partir da Eq. 1, com pesos w_i proporcionais ao tamanho dos *dataset*s $|D_i|$ de cada cliente.

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

Neste caso w_i é o peso atribuído a cada cliente com base no tamanho do seu conjunto de dados D_i, e θ_i são os parâmetros locais treinados por cada cliente calculados a partir da Eq. 2. Assim, o FedAvg permite que o modelo global seja ajustado com base na contribuição proporcional de cada cliente, considerando o tamanho de seus dados locais.

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

3.2. Funcionamento do SCOPE-FL

O princípio fundamental por trás dessa abordagem é que a relevância de cada cliente para o treinamento do modelo global pode ser diretamente relacionada ao tamanho de seu conjunto de dados local. Assim, a ideia é atribuir um peso maior aos clientes que possuem conjuntos de dados maiores, uma vez que eles têm uma maior capacidade de influenciar a qualidade do modelo global.

Nesta abordagem busca-se selecionar os clientes com maior relevância para o treinamento do modelo global, utilizando dois critérios principais: o **tamanho do *dataset*** local e a **entropia** dos dados do cliente após a primeira etapa de treinamento. A relevância de cada cliente é determinada pela combinação desses dois fatores, sendo que um cliente com um conjunto de dados maior e um valor de entropia mais elevado, calculado conforme a Eq. 3, é considerado mais relevante para contribuir com a atualização do modelo global. A entropia dos dados $H(x)$, onde $P(x)$ denota a probabilidade de observar um valor específico x no conjunto de dados, e $\log$ é o logaritmo natural.

$$H(X) = -\sum_{x} P(x) \log P(x) \tag{3}$$

Para a fórmula dos pesos, cada cliente w_i pode ser modelada como uma função que considera, a partir de uma filtragem prévia dos dados, o ranking dos 40% do **tamanho do *dataset*** $|D_i|$ quanto à entropia calculada por Eq. 3.

Eq.4 apresentada calcula a pontuação de relevância R_i para cada cliente i, considerando dois fatores principais: o tamanho do *dataset* do cliente D_i e a entropia dos dados H_i após a primeira etapa de treinamento. O tamanho do *dataset* D_i representa o número de amostras dos dados do cliente i, e D_{max} é o maior tamanho de *dataset* entre os clientes selecionados (os 40% com os maiores *dataset*s). Já H_i é a entropia dos dados do cliente i, enquanto H_{max} é o maior valor de entropia entre os clientes selecionados. Os pesos α e β são atribuídos ao tamanho do *dataset* e à entropia, respectivamente, permitindo ajustar a importância de cada fator na pontuação final.

$$R_i = \alpha \cdot \frac{D_i}{D_{\text{max}}} + \beta \cdot \frac{H_i}{H_{\text{max}}} \tag{4}$$

A pontuação R_i é uma combinação linear da normalização do tamanho do *dataset* e da entropia, sendo feita em relação aos valores máximos dentro do conjunto de clientes selecionados. Isso permite selecionar clientes com *dataset*s grandes e/ou dados com alta entropia, de acordo com a importância atribuída a cada fator pelos pesos α e β. Dessa forma, a equação ajuda a identificar os clientes mais relevantes para a contribuição no treinamento do modelo global, equilibrando tanto o volume de dados quanto a diversidade ou complexidade dos dados do cliente.

4. Avaliação

Nesta seção, são apresentados os resultados da comparação entre o estado da arte e o SCOPE-FL, considerando os indicadores de perda, acurácia e tempo, visando melhorar a performance da simulação de FL. O estudo foi realizado utilizando o framework PFLlib [Zhang et al. 2023] em um servidor, com o conjunto de dados público MNIST. A arquitetura do modelo é uma *Convolutional Neural Network* (CNN) com duas camadas convolucionais de filtros 5x5, seguidas de max-pooling 2x2 após cada camada. Foram utilizados 100 clientes em 50 rodadas, com 10% deles selecionados para o treino.

A Figura 1 apresenta a medição de acurácia para três abordagens comparadas: a Loss baseada no trabalho de Souza *et al.* [de Souza et al. 2023], a abordagem aleatória (Random) e a proposta deste trabalho, o SCOPE-FL. Neste cenário, com apenas 10% dos clientes selecionados, observa-se que, com poucos ajustes e um número mínimo de

clientes, o SCOPE-FL atinge mais de 60% de acurácia em 12 rodadas e continua progredindo até cerca de 80% de acurácia após 22 rodadas. Isso demonstra que o SCOPE-FL possui um bom desempenho, alcançando resultados elevados com poucas interações. Em contraste, o método de Loss apresenta uma progressão mais lenta, atingindo uma convergência máxima média de 64%. O método Random, por sua vez, apresenta resultados semelhantes aos do SCOPE-FL, mas com uma convergência menos suave, alcançando apenas 56% de acurácia após 12 rodadas, embora ainda progrida para valores próximos aos do SCOPE-FL nas últimas rodadas.

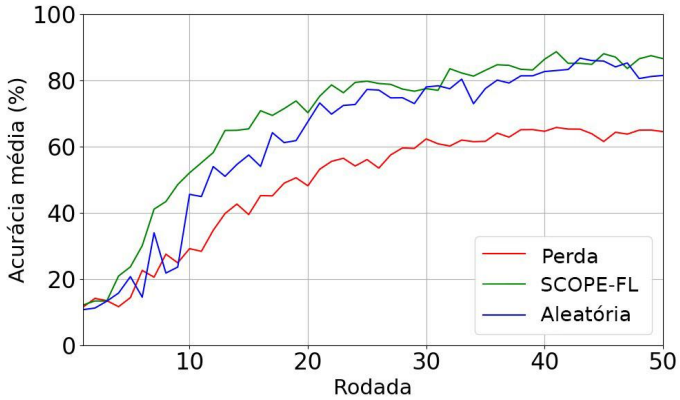

Figura 1. Medição de Acurácia

A Figura 2 mostra a medição de perda dos modelos, evidenciando como a seleção de clientes com melhores preferências de *dataset* impacta o aprendizado. Com mais dados e maior importância deles, os modelos nos clientes conseguem aprender de forma mais eficaz. O SCOPE-FL apresentou os melhores resultados, alcançando valores de perda abaixo de 1 após cerca de 15 rodadas. Em contrapartida, o método de Loss não teve o mesmo desempenho, não conseguindo reduzir a perda abaixo de 1, mostrando um gargalo de aprendizado por volta da 30ª rodada. O método random, embora com desempenho semelhante ao do SCOPE-FL, não superou o modelo proposto neste artigo, alcançando resultados melhores apenas após cerca de 40 rodadas, quando ambos já estavam em fase de convergência.

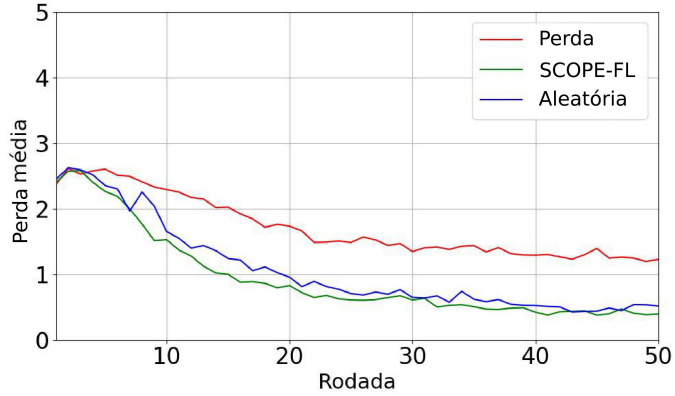

Figura 2. Medição de Perda

5. Conclusão

Este trabalho propôs uma solução para melhorar a seleção de clientes em FL por meio do algoritmo SCOPE-FL, que utiliza a entropia dos dados e o tamanho dos *dataset*s locais para otimizar a escolha dos clientes mais relevantes para o treinamento do modelo global. O SCOPE-FL demonstrou ser eficaz na melhoria da precisão e eficiência do treinamento, superando métodos tradicionais, como Loss e Random, especialmente em termos de acurácia e velocidade de convergência. As simulações realizadas com o MNIST mostraram que o SCOPE-FL obteve mais de 60% de acurácia após 12 rodadas.

Como trabalho futuro, a implementação de métodos adaptativos de seleção de clientes com base em características dinâmicas de dados e o aprimoramento da abordagem SCOPE-FL podem aumentar ainda mais a robustez e a escalabilidade do FL. A exploração de diferentes cenários de dados não independentes e idênticos (non-IID) pode expandir o alcance e a aplicabilidade da metodologia, proporcionando uma solução mais eficiente e privada para sistemas de FL em dispositivos de borda. Assim, utilizando métodos como a distribuição patológica, em que a distribuição é ajustada para cenários mais desafiadores, implica no uso de *dataset*s ainda mais complexos, aumentando a dificuldade do modelo.

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.

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.

Deng, Y., Lyu, F., Ren, J., Wu, H., Zhou, Y., Zhang, Y., and Shen, X. (2021). Auction: Automated and quality-aware client selection framework for efficient federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(8):1996–2009.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Orlandi, F. C., Dos Anjos, J. C., Santana, J. F. d. P., Leithardt, V. R., and Geyer, C. F. (2023). Entropy to mitigate non-iid data problem on federated learning for the edge intelligence environment. *IEEE Access*.

Pires, I. M., Marques, G., Garcia, N. M., Flórez-Revuelta, F., Canavarro Teixeira, M., Zdravevski, E., Spinsante, S., and Coimbra, M. (2020). Pattern recognition techniques for the identification of activities of daily living using a mobile device accelerometer. *Electronics*, 9(3):509.

Zhang, J., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., and Guan, H. (2023). Fedala: Adaptive local aggregation for personalized federated learning. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 11237–11244.