FedSeleKDistill: Empoderando a Escolha de Clientes com a Destilação do Conhecimento para Aprendizado Federado em Dados Não-IID

Aissa H. Mohamed[1], **Allan M. de Souza**[1], **Joahannes B. D. da Costa**[1], **Leandro A. Villas**[1], **Julio C. Dos Reis**[1]

[1]Instituto de Computação (UNICAMP), Campinas – São Paulo – Brasil

a265189@dac.unicamp.br, {allanms, jbdc, lvillas, jreis}@unicamp.br

Abstract. *Federated Learning (FL) is a distributed approach in which multiple devices collaborate to train a shared global model (GM). During its training, client devices must communicate their gradients to the central server to update the GM weights. This incurs significant communication costs (bandwidth utilization and the number of messages exchanged). The heterogeneous nature of clients' local datasets poses an extra challenge to the training. In this sense, we introduce FedSeleKDistill,* ***Fed****erated* ***Sele****ction and* ***K****nowledge* ***Distill****ation* ***A****lgorithm, to decrease the overall communication costs. FedSeleKDistill is an innovative combination of: (i) client selection, and (ii) knowledge distillation approaches with three main objectives: (i) reducing the number of devices training at every round; (ii) decreasing the number of rounds until convergence; and (iii) mitigating the effect of client's heterogeneous data on the GM effectiveness. Our experimental evaluations on the well-known dataset MNIST demonstrate that FedSeleKDistill is highly efficient in training the GM until convergence. FedSeleKDistill reaches a higher accuracy score and faster convergence than state-of-the-art models. Our results also show higher performance when analyzing the accuracy scores on the client's local datasets.*

Resumo. *FL é uma abordagem distribuída na qual múltiplos dispositivos colaboram para treinar um modelo global compartilhado. Durante seu treinamento, os dispositivos clientes devem comunicar seus gradientes ao servidor central para atualizar os pesos do modelo global. Isso acarreta custos significativos de comunicação (utilização de largura de banda e o número de mensagens trocadas). A natureza heterogênea dos conjuntos de dados dos clientes representa um desafio adicional. Nesse sentido, introduzimos o FedSeleKDistill,* ***Fed****erated* ***Sele****ction and* ***K****nowledge* ***Distill****ation* ***A****lgorithm, para reduzir os custos de comunicação globais. O FedSeleKDistill é uma combinação inovadora de: (i) seleção de clientes e (ii) abordagens de destilação de conhecimento com três objetivos principais: (i) reduzir o número de dispositivos treinando em cada rodada; (ii) diminuir o número de rodadas para atingir a convergência; e (iii) mitigar o efeito dos dados heterogêneos do cliente na eficácia do modelo global. Nossas avaliações experimentais em conjunto de dados MNIST demonstram que o FedSeleKDistill é altamente eficiente no treinamento do modelo global até a convergência. O FedSeleKDistill alcança uma pontuação de precisão mais alta e uma convergência mais rápida do que os modelos baseados em estado da arte. Nossos resultados também mostram um desempenho superior ao analisar as pontuações de precisão nos conjuntos de dados locais do cliente.*

1. Introdução

Nos últimos anos, tem havido um considerável desenvolvimento na criação de uma variedade de dispositivos computacionais, tais como *smartphones* e dispositivos de Internet das Coisas [de Souza et al. 2023]. Esses equipamentos geram vastos conjuntos de dados, que por sua vez têm sido utilizados pelos pesquisadores para alcançar avanços significativos no treinamento de modelos de Aprendizado Profundo, do inglês *Deep Learning (DL)*, abrangendo áreas como reconhecimento de objetos/padrões, processamento de linguagem natural e agentes autônomos, como robôs e veículos automotores. Contudo, esse progresso tem sido acompanhado por uma crescente demanda por dados e recursos computacionais, o que suscita questões sobre ética e privacidade.

O Aprendizado Federado, do inglês *Federated Learning (FL)*, emergiu como uma abordagem inovadora para enfrentar esses desafios [Li et al. 2020]. Em suma, o FL envolve o treinamento de modelos de DL usando dados distribuídos em dispositivos individuais, chamados de clientes. Esses clientes treinam um modelo compartilhado enquanto garantem a privacidade de seus dados locais. Dentro do FL, o processo de treinamento começa com os clientes baixando os pesos do modelo global do servidor central. Em seguida, eles treinam esse modelo usando seus dados locais e transmitem seus gradientes locais para o servidor central. O servidor central atualiza os pesos do modelo global por meio de uma média e retorna os parâmetros do modelo recém-atualizados para os clientes participantes. Então, uma nova rodada de treinamento começa. Esse processo iterativo continua até que o modelo global alcance um valor de perda razoável. Nesse caso, a literatura diz que o modelo atingiu a convergência.

No contexto do FL, surgem diversos desafios relacionados à comunicação entre os dispositivos clientes e o servidor central ([Shahid et al. 2021], [de Souza et al. 2024]). Os dispositivos conectados à rede precisam constantemente compartilhar suas atualizações, o que pode resultar em um gargalo na comunicação [Mothukuri et al. 2021]. Além disso, os dispositivos participantes nem sempre contam com uma conexão de comunicação adequada ou confiável [Lim et al. 2020]. Devido à largura de banda limitada e aos recursos energéticos dos dispositivos clientes, as rodadas de comunicação podem ser demoradas [Li et al. 2020]. Um outro desafio é o fato que os dados dos clientes que participam do processo de treinamento podem não seguir uma distribuição independente e identicamente distribuída (IID) [Li et al. 2018]. Esses dados refletem informações específicas de cada cliente, como padrões de uso, preferências do usuário e informações do ambiente local [Shahid et al. 2021]. Portanto, o conjunto de dados de um cliente pode não ser representativo da distribuição de dados de toda a população [McMahan et al. 2017].

Assim, torna-se crucial otimizar a eficiência da comunicação em um ambiente de FL [Mothukuri et al. 2021]. Um algoritmo de treinamento distribuído eficiente em comunicação para FL precisa atender aos seguintes requisitos [Sattler et al. 2020]: (*i*) deve reduzir as comunicações entre os clientes e o servidor, (*ii*) deve ser robusto para dados não-IID, tamanhos de lote pequenos e dados desbalanceados, e (*iii*) deve ser robusto para grandes números de dispositivos e participação parcial do cliente.

Diante desses desafios, é essencial realizar a comunicação em um ambiente de FL da maneira mais eficiente possível [Mothukuri et al. 2021]. Nesse sentido, propomos um novo algoritmo de treinamento, chamado FedSeleKDistill (**Fed**erated **Sele**ction and

Knowledge **Distill**ation **A**lgorithm), que aborda com sucesso os desafios mencionados para o treinamento distribuído em FL. Nossa solução proposta é eficiente em relação ao número de rodadas necessárias para alcançar a convergência e eficaz em cada rodada para selecionar os clientes para o treinamento. Nossos experimentos empíricos em um conjunto de dados MNIST mostram a superioridade de FedSeleKDistill em comparação com o *Power-Of-Choice* (o POC) [Cho et al. 2020] e o FedAvg [McMahan et al. 2017].

A próxima seção, 2, oferece uma análise concisa da literatura existente sobre abordagens de FL para lidar com desafios relacionados à comunicação. A seção 3 apresenta o nosso algoritmo de treinamento chamado FedSeleKDistill. A seção 4 avalia o desempenho da solução proposta, enquanto a seção 5 discute os resultados obtidos e encerra este artigo com considerações finais.

2. Trabalhos Relacionados

Os pesquisadores têm adotado diversas abordagens para enfrentar desafios específicos no contexto do FL. Essas abordagens podem ser categorizadas em várias classes, conforme proposto por [Shahid et al. 2021]. Nesta seção, são apresentados alguns estudos recentes em FL, organizados de acordo com essas categorias. Por exemplo, Federated Averaging (FedAvg) [McMahan et al. 2017] reduz o número de comunicações aumentando o número de iterações por cliente por rodada antes de enviar os gradientes locais dos clientes para o servidor central atualizar o modelo global. No entanto, para garantir a convergência, FedAvg supõe que a distribuição dos dados entre os clientes seja *IID*.

Outra abordagem considera a compressão das mensagens entre o servidor central e os clientes. Podemos citar as técnicas de quantização ([Lin et al. 2017, Amiri et al. 2020, Bernstein et al. 2018]) e esparsificação ([Sattler et al. 2020, Rothchild et al. 2020, Ström 2015]). No entanto, nem todos esses métodos de compressão garantem a convergência do modelo global [Shahid et al. 2021].

Outra abordagem é a seleção de clientes para treinamento a cada rodada. O *Power-Of-Choice* (POC) [Cho et al. 2020] é um *framework* de seleção eficiente em comunicação e computação que tende a selecionar clientes com perda local mais alta. Avaliações experimentais demonstraram que o POC alcança uma convergência mais rápida, aumenta a eficiência de comunicação e reduz os custos gerais de comunicação. Similarmente, o FOLB [Nguyen et al. 2020] realiza uma amostragem inteligente de clientes em cada rodada de treinamento. Como mencionado pelos autores, clientes específicos proporcionam melhorias mais significativas ao modelo global do que outros durante o treinamento.

Em [Wang et al. 2020], os autores utilizam *Deep Reinforcement Learning (DRL)* para treinar um agente chamado *FAVOR* para selecionar um conjunto ótimo de clientes em cada rodada de treinamento. O *FAVOR* escolhe de forma inteligente os dispositivos clientes para equilibrar o viés introduzido por dados não-IID e acelerar a convergência. No entanto, o *FAVOR* precisa de novo treinamento se exposto a novas condições de ambiente. Além disso, não conseguimos entender completamente os critérios de *FAVOR* para selecionar os clientes ideais a cada rodada, pois ele é uma caixa-preta não interpretável. Em [Mohamed et al. 2023b], os autores propõem o *framework* CCSF que agrupa os clientes em grupos homogêneos. Em seguida, usando uma das três estratégias de seleção, o CCSF cria um subconjunto de clientes para treinamento a cada rodada. Resultados experimentais mostraram que o CCSF treina um modelo global até a convergência mais

rapidamente do que o FedAvg. No entanto, uma limitação dessa solução é que o CCSF requer que todos os clientes sejam agrupados em *clusters* homogêneos.

Uma abordagem cada vez mais adotada em trabalhos recentes é a Destilação de Conhecimento (*Knowledge Distillation (KD)*). De acordo com [Mora et al. 2022], a KD pode ser empregada para dois objetivos. O primeiro é permitir que os clientes participantes selecionem diferentes arquiteturas de modelo (heterogeneidade de modelo). O segundo é mitigar o impacto da heterogeneidade de dados no desempenho do modelo global. Como o foco do presente trabalho é abordar o problema de dados não-IID, apresentamos alguns trabalhos recentes com o segundo objetivo.

[Mora et al. 2022] distinguem entre estratégias do lado do servidor que refinam a agregação do FedAvg com uma fase de destilação e técnicas do lado do cliente que destilam localmente o conhecimento global para lidar com a deriva do cliente. Como solução baseada em KD do lado do servidor, o FedDF [Lin et al. 2020] usa um conjunto de dados *proxy* para ajustar finamente o modelo global imitando a saída do modelo de conjunto dos clientes. Uma limitação é que o FedDF supõe que o servidor central e os clientes compartilham um conjunto de dados comum (conjunto de dados *proxy*).

No caso de ausência de dados comuns, [Zhang et al. 2022] propõem FedFTG que modela o espaço de entrada dos modelos locais por meio de um gerador auxiliar no servidor central, e então gera dados comuns para transferir o conhecimento nos modelos locais para o modelo global, a fim de melhorar o desempenho.

Por outro lado, o FedGKD [Yao et al. 2021] funde o conhecimento de modelos globais históricos para o treinamento local para mitigar o problema da ”deriva do cliente”. À medida que os clientes realizam atualizações locais em dados heterogêneos, seus modelos locais se afastam. Como resultado, o modelo global pode sofrer *overfitting* nos dados locais de alguns clientes. O FedGKD usa a KD do lado do cliente para orientar o treinamento do modelo local pelos professores globais (modelos globais passados), onde cada cliente aprende o conhecimento global dos modelos globais passados via técnicas de destilação adaptativa de conhecimento.

Os autores [Lee et al. 2022] observam que, ao ser treinado nos dados locais do cliente, o modelo global esquece o conhecimento das rodadas anteriores. Com base nesses achados, os autores propõem o *Federated Not-True Distillation* (o FedNTD) para lidar com o problema de esquecimento. O FedNTD adota a destilação local do conhecimento global para mitigar o esquecimento entre as rodadas subsequentes e amenizar a prejudicialidade da heterogeneidade de dados.

De acordo com [He et al. 2022b], quando o modelo global não convergiu completamente e não aprendeu completamente a distribuição dos dados locais dos clientes, o desempenho do modelo global pode ser melhor ou pior para algumas classes. [He et al. 2022b] propõem um mecanismo de Auto-Destilação Adaptativa por Classe (o FedCAD) para lidar com esse problema. Sob essa solução, o FedCAD utiliza termos adaptativos por classe para suavizar a influência da perda de destilação de acordo com o desempenho do modelo global em cada classe. Utilizando uma abordagem semelhante, [He et al. 2022a] propõem um método de Auto-Destilação Seletiva (o FedSSD), que impõe restrições adaptativas nas atualizações locais ao auto-destilar o conhecimento do modelo global e ponderá-lo seletivamente avaliando a credibilidade tanto a nível de

classe quanto de amostra. Avaliações experimentais mostram que o FedSSD alcança melhor generalização e robustez em menos rodadas de comunicação do que outros métodos de FL de última geração.

3. Proposta

Esta seção apresenta o FedSeleKDistill, que aborda três problemas críticos em FL: *(i)* os dados não-IID; *(ii)* o alto *overhead* de comunicação; e *(iii)* a taxa limitada de participação do cliente. O primeiro é tratado por meio da destilação de conhecimento. Os dois últimos problemas são resolvidos pela estratégia de seleção de cliente enviesada para aumentar a velocidade de convergência do modelo.

3.1. Enunciado do problema

Vamos considerar uma configuração de FL entre dispositivos com um total de N clientes, onde o cliente $n \in N$ possui um conjunto de dados local D_n consistindo de $|D_n|$ amostras de dados. Nosso objetivo é que os clientes treinem localmente um modelo global com pesos iniciais W_{init}, compartilhem seus pesos locais w_n com o servidor central, que é responsável por agregar os pesos de cada cliente, e coletivamente encontrar o parâmetro do modelo w^* que minimize:

$$F(w) = \frac{1}{\sum_{n \in N} |D_n|} \sum_{n \in N} \sum_{d_n \in D_n} f(w_n, d_n) \tag{1}$$

O que pode ser escrito como:

$$F(w) = \sum_{n \in N} p_n F_n(w_n) \tag{2}$$

onde $f(w_n, d_n)$ é a função de perda composta para a amostra $d_n \in D_n$ e os pesos w_n do cliente n. O termo $p_n = \frac{|D_n|}{\sum_{n \in N} |D_n|}$ é a fração de dados do cliente n, e $F_n(w_n) = \frac{1}{|D_n|} \sum_{d \in D_n} f(w, d)$ é a função objetivo local do cliente n.

3.2. Descrição de FedSeleKDistill

Os objetivos do FedSeleKDistill são (*i*) treinar um modelo global até a minimização da função de perda (1) para o mínimo local, considerando as restrições no ambiente de FL (*por exemplo*, dados não-IID, participação limitada do cliente e recursos computacionais disponíveis); (*ii*) alcançar a convergência com o menor número possível de rodadas de comunicação; e (*iii*) ter uma comunicação eficiente cliente-servidor. O FedSeleKDistill é baseado em uma estratégia de seleção de cliente enviesada que torna o treinamento eficiente em relação ao número de rodadas necessárias para treinar o modelo global até a convergência. Para mitigar o efeito negativo da heterogeneidade de dados entre os clientes de FL, o FedSeleKDistill utiliza a destilação de conhecimento KD local.

3.2.1. Destilação de Conhecimento

A destilação de conhecimento é uma técnica de transferência de aprendizado em que um modelo (pre-treinado) mais complexo e robusto, chamado de professor, é usado para

treinar um modelo menor e mais simples, conhecido como aluno ([Bucila et al. 2006], [Hinton et al. 2015]). O objetivo principal da destilação de conhecimento é transferir o conhecimento do professor para o aluno, permitindo que o aluno adquira uma representação mais compacta e eficiente dos padrões aprendidos pelo professor. Isso é especialmente útil quando o professor é um modelo grande e computacionalmente caro, e o aluno precisa ser implantado em dispositivos com recursos limitados, como dispositivos móveis ou sistemas embarcados.

No contexto de FL, a regularização baseada em KD local tem como objetivo reduzir de forma eficaz a influência de dados não independentes e identicamente distribuídos (não-IID) [Mora et al. 2022]. A função objetivo local (1) no dispositivo do cliente torna-se uma combinação linear entre a perda de entropia cruzada e uma perda baseada em KD, que avalia a diferença entre a saída do modelo global (modelo professor) e a saída do modelo local (modelo aluno) nos dados locais, utilizando a divergência de Kullback-Leibler, conforme descrito a seguir:

$$L_{local} = (1 - \lambda) \cdot L_{CE} + \lambda \cdot L_{KD} \tag{3}$$

onde λ é o peso da divergência de Kullback-Leibler, com um valor entre 0 e 1 que determina a contribuição do termo de perda KL na perda local do cliente L_{local}, L_{CE} é a perda de entropia cruzada e L_{KD} é a perda baseada em KD.

Usando as notações na Figura 1, a função 3 para o cliente k torna-se:

$$f(w_k, d_k) = (1 - \lambda) \left(-\frac{1}{|D_k|} \sum_{(x_i, y_i) \in D_k} y_i \cdot \log(\hat{y}i, wk, t + 1) \right) \quad + \lambda \text{KL} \left(P_{w_t}(y_i), ||, P_{w_{k,t+1}}(y_i) \right) \tag{4}$$

onde D_k são os dados locais do cliente k, (x_i, y_i) é a amostra de dados i. w_t representa os pesos do modelo global (modelo professor) na rodada t, $w_{k,t+1}$ representa o modelo local (modelo aluno) treinado nos dados locais D_k na rodada t+1.

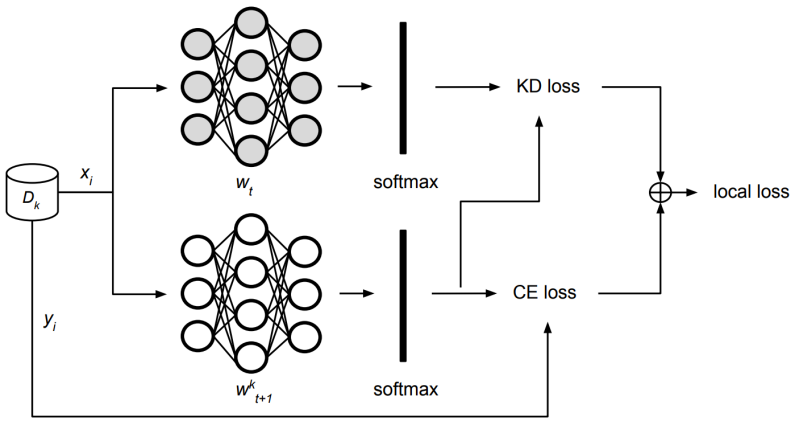

Figura 1. O *framework* do mecanismo de destilação de conhecimento local no lado do cliente utilizado [Mora et al. 2022].

A Figura 1 descreve o *framework* do mecanismo de destilação de conhecimento local no lado do cliente utilizado pelo FedSeleKDistill baseado em [Mora et al. 2022].

Depois de ser selecionado para treinar, o dispositivo do cliente cria uma cópia do modelo global. O modelo original atua como o professor e transfere seu conhecimento para a cópia (o aluno) por meio da destilação usando os dados locais do cliente. O modelo aluno é o modelo local treinado com a função objetivo local 4. Além disso, esse processo de destilação de conhecimento preserva a privacidade dos dados dos clientes.

3.2.2. Power-Of-Choice

A segunda ideia-chave é inspirada no método *Power-Of-Choice* (o POC) [Cho et al. 2020] para selecionar os clientes para o treinamento. Segundo os autores, fazer uma seleção de clientes enviesada considerando os clientes com a maior perda local $f_n(w)$ pode melhorar a convergência do modelo global. Por exemplo, consideremos um cenário com dois clientes, um dos quais apresenta uma alta perda local e o outro uma perda local significativamente menor. Em uma determinada rodada durante a fase de treinamento, o modelo global obtém um aprendizado mais significativo, concentrando-se nos dados locais do cliente com a pior perda.

Matematicamente, o cliente com a pior perda local possui um gradiente de magnitude maior do que o cliente com a menor perda local. O termo $\frac{\partial L}{\partial w}$ representa o gradiente da função de perda em relação aos pesos do modelo global. O magnitude do gradiente, $\left\|\frac{\partial L}{\partial w}\right\|$, refere-se ao tamanho do gradiente. Esse termo aumenta quando o valor L, representando a perda local, é maior. Portanto, se atualizarmos os pesos do modelo global escolhendo o gradiente com a maior magnitude, o modelo global fará um progresso maior na próxima etapa de atualização em direção ao mínimo local da função de perda 1. No entanto, quanto mais enviesada for a seleção de clientes considerando os clientes com a maior perda para obter uma convergência mais rápida, maior será o risco de haver uma lacuna não desaparecente entre os pesos ótimos verdadeiros e os pesos na convergência para o modelo global com essa estratégia [Cho et al. 2020]. Este é o *trade-off* entre a velocidade de convergência e o viés de solução mencionado pelos autores [Cho et al. 2020].

3.2.3. Descrição do FedSeleKDistill

O FedSeleKDistill combina o algoritmo de treinamento proposto no *framework* POC [Cho et al. 2020] com a destilação de conhecimento local. Portanto, o FedSeleKDistill adota duas estratégias. A primeira é uma estratégia de seleção enviesada de clientes para treinar o modelo global até a convergência com poucas rodadas em comparação com o FedAvg. A segunda abordagem mitiga o efeito negativo da heterogeneidade de dados nos dispositivos dos clientes. Especialmente quando o número de *iterações* (as épocas) por cliente por rodada é alto, o risco do modelo global sofrer *overfitting* nos dados locais aumenta. Além disso, a KD local tem a vantagem adicional de evitar que o modelo global esqueça o conhecimento adquirido em rodadas anteriores a partir dos dados de clientes passados quando treinado com dados locais de um novo cliente.

Com base nos trabalhos propostos em [Cho et al. 2020], apenas o algoritmo de treinamento local precisa ser modificado conforme descrito no Algoritmo 1. Na solução proposta, o valor de λ é inicialmente selecionado, entre 0 e 1, e permanece constante durante as rodadas de treinamento. Geralmente, o valor de λ é definido como 0.5.

Algoritmo 1: Dispositivo do Cliente

▷ Quando o cliente n recebe uma solicitação de treinamento

LocalTrain *(w)*:

  for *época* $\in \{1, 2, \ldots, E\}$ **do**

    foreach *batch* $\in D_n$ **do**

      ▷ ajusta o modelo com dados locais

      $w_n \leftarrow (1 - \lambda)L_{CE}(w_t, batch) + \lambda L_{KD}(w_t, batch)$

    ▷ envia modelo treinado para o servidor

    return w_n

4. Avaliação

Esta seção descreve a metodologia experimental e as métricas utilizadas para avaliar a eficiência do FedSeleKDistill.

4.1. Descrição do Conjunto de Dados

Selecionamos o conjunto de dados MNIST para os experimentos. Esse conjunto contém um total de 70.000 imagens em tons de cinza, divididas em um conjunto de treinamento com 60.000 imagens e um conjunto de teste com 10.000 imagens. Cada imagem de 28x28 *pixels* retrata um dígito manuscrito (0 a 9). Os detalhes sobre os dados podem ser encontrados em [LeCun et al. 1998]. A tarefa é um problema de classificação, onde treinamos um modelo de aprendizado profundo para classificar as imagens entre 10 classes (0 a 9). Para o experimento, definimos o número total de clientes FL para 30. Dividimos as 60.000 imagens igualmente entre todos os clientes. Cada conjunto de dados local contém 70% de uma classe, e os 30% restantes são uma distribuição uniforme das outras classes. O conjunto de dados locais criados para serem distribuídos a todos os clientes FL seguiram a metodologia utilizada em [Wang et al. 2020]. O modelo global selecionado para aprender o conjunto de dados MNIST é uma rede neural profunda sequencial composta por 2 camadas de convolução e *pooling*, seguidas por uma camada de achatamento, regularização de *dropout* e uma camada densa com ativação *softmax* para prever classes.

Nesses experimentos, avaliamos o modelo global em cada rodada de treinamento FL em um conjunto de testes localizado no servidor central. O objetivo na tarefa de classificação é alcançar uma pontuação de precisão no conjunto de testes que seja o mais próxima possível das pontuações de precisão de 98.94% no ambiente centralizado.

4.2. Métricas Avaliadas

Nesta análise, avaliamos nosso algoritmo de treino FedSeleKDistill em comparação com o POC e o FedAvg usando a pontuação de precisão ao longo das rodadas de FL como a métrica. Quanto menos rodadas forem necessárias para atingir a precisão máxima, mais rápida será a convergência da solução.

4.3. Resultados Experimentais

Em nossos experimentos, definimos o número total de clientes em FL como 30. A taxa de participação, definida como o número de clientes disponíveis para treinamento em cada rodada, é definida como 10 ou 33.33%. O número de clientes selecionados para

treinamento em cada rodada é definido como 5 ou 50% da taxa de participação. Durante os experimentos, avaliamos o FedSeleKDistill em comparação com o POC e o FedAVG aumentando o número de épocas por cliente por rodada. Tabela 1 resume os detalhes sobre os experimentos.

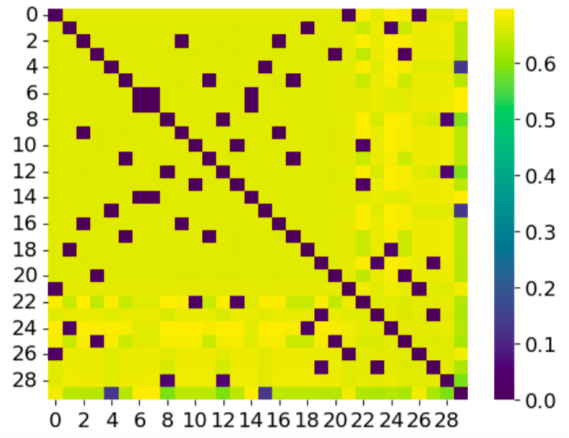

Figura 2. Os testes de Kolmogorov-Smirnov são realizados em cada par de clientes presentes no conjunto de dados (MNIST). Quanto maior o valor, de 0 a 1, maior o nível de heterogeneidade entre os conjuntos de dados de dois clientes.

A Figura 2, representando a matriz de teste de Kolmogorov-Smirnov entre todos os pares de clientes, mostra que o nível de heterogeneidade dos dados locais de clientes é relativamente alto.

Tabela 1. Parâmetros de Treinamento de FL

Parâmetros	**Valor**
Número total de clientes	30
Taxa de participação (clientes/rodada)	10
Taxa de seleção (clientes/rodada)	5
Número de iterações (*epochs*)	3, 4, 5

A Figura 3 mostra a pontuação de precisão do modelo global em cada rodada, no conjunto de testes localizado no servidor central.

O FedAVG é o algoritmo de treinamento menos eficiente em comparação com o POC e FedSeleKDistill nos 3 experimentos. Por exemplo, o FedAVG alcança uma pontuação de precisão de 80% após mais de 40 rodadas, enquanto o POC e FedSeleK-Distill atingem 80% após 18 e 13 rodadas respetivamente, com 3 *epochs* (iterações). Isso mostra a importância de adotar uma estratégia de seleção de cliente para treinamento a fim de alcançar uma convergência mais rápida. Também, o modelo global atinge uma precisão superior a 90% após 21 rodadas com o FedSeleKDistill, contre 36 com o POC (gráfico ezquerdo). Com 4 *epochs* (gráfico do meio), o desempenho de FedSeleKDistill é mais rápido que o POC, atingindo mais que 94% de precisão em 33 rodadas enquanto o POC necessita 50 rodadas de treinamento. Considerando 5 *epochs* (gráfico direito), o FedSeleKDistill permite o modelo global atingir uma precisão de 94.45% em 27 rodadas, enquanto o POC atinge ao máximo 91% de precisão com 30 rodadas de treinamento.

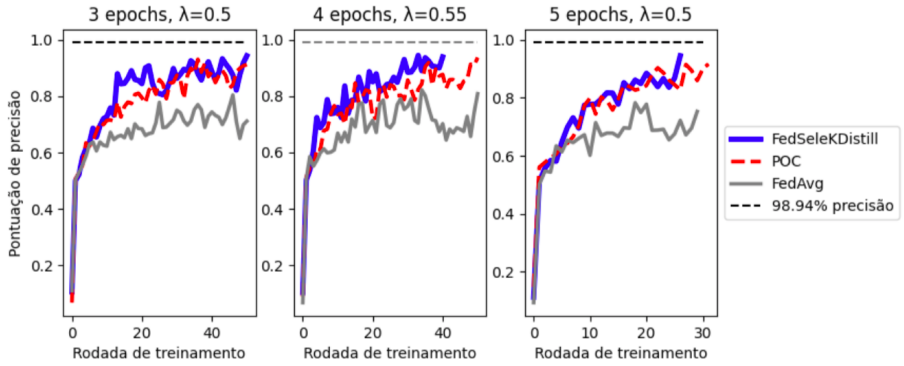

Figura 3. Pontuação de precisão no conjunto de teste localizado no servidor central (MNIST), a cada rodada de treinamento, com 3, 4 e 5 *epochs* de treinamento por cliente selecionado por rodada.

Quanto mais aumentamos o número de *epochs*, maior é a diferença em termos de rodadas necessárias até a convergência entre o nosso algoritmo de treinamento e o POC.

Estes experimentos mostram, em primeiro lugar, que nosso algoritmo de treinamento, combinando o POC e a destilação do conhecimento, treina com sucesso o modelo global até a convergência. Segundo, a destilação do conhecimento permite que o modelo global aprenda de forma mais eficiente a distribuição dos dados locais, como pode ser observado na Figura 4.

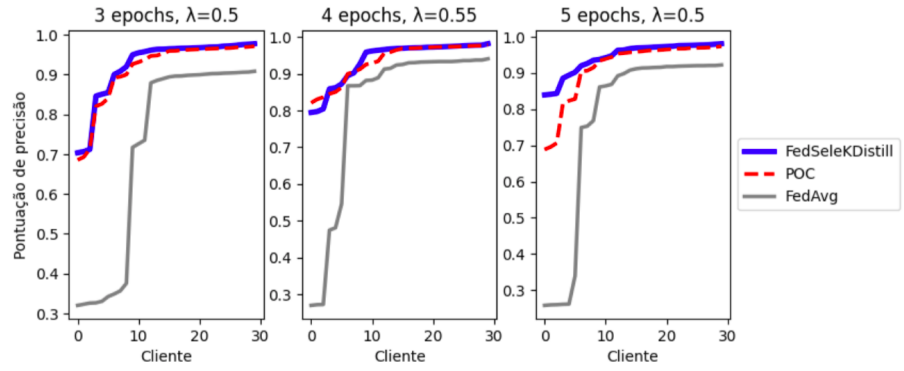

Figura 4. Pontuação de precisão, depois do treinamento do modelo global, nos dados locais de clientes (MNIST).

A Figura 4 mostra a pontuação de precisão do modelo global avaliado nos conjuntos de dados locais dos clientes, após 50 rodadas com 3 *epochs*. 40 rodadas de treinamento no caso de FedSeleKDistill, 50 no caso do POC e o FedAVG com 4 *epochs*. E finalmente, após 27 rodadas no caso de FedSeleKDistill, e 30 no caso do POC e o FedAvg com 5 *epochs*. Como podemos ver, a KD permite que o FedSeleKDistill aprenda melhor e de maneira mais uniforme as distribuições de dados locais do que o POC. Para cada cliente, o modelo global alcançou uma pontuação de precisão mais alta em cada conjunto de dados local quando treinado usando o FedSeleKDistill em comparação com o POC e o FedAVG (exceto no caso de 4 *epochs* onde o FedSeleKDistill não é superior à POC em todos os clientes).

É importante mencionar que, para alcançar esses resultados, o valor de λ teve que

ser alterado ao mudar o número de iterações por cliente por rodada.

4.4. Diminuindo o nível de heterogeneidade

Nesta seção, utilizamos um conjunto de dados de reconhecimento de atividade humana (*Human Activity Recognition*, HAR), construído a partir das gravações de 30 participantes do estudo realizando atividades da vida diária enquanto carregavam um *smartphone* montado na cintura com sensores inerciais embutidos. Os detalhes sobre os dados podem ser encontrados em [Kaggle]. Em nosso experimento, cada cliente representa um participante (21 em total). O conjunto de dados representando os 9 outros participantes serve como dados de teste no servidor central para avaliar o desempenho do modelo global em cada rodada de treinamento.

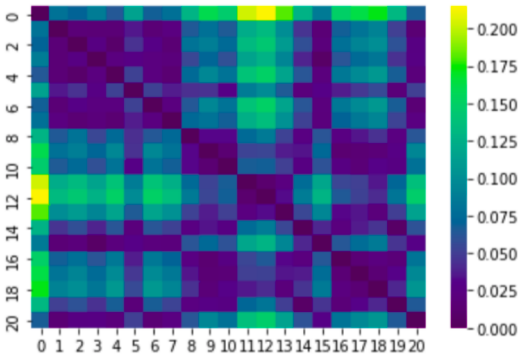

Figura 5. Os testes de Kolmogorov-Smirnov são realizados em cada par de clientes presentes no conjunto de dados (HAR). Quanto maior o valor, de 0 a 1, maior o nível de heterogeneidade entre os conjuntos de dados de dois clientes [Mohamed et al. 2023a].

Embora os 21 conjuntos de dados locais dos clientes pareçam ser relativamente heterogêneos (figura 5), ao avaliar o modelo global sobre os conjuntos de dados locais dos clientes durante o treinamento, a pontuação de precisão é uniforma entre os 21 clientes a cada rodada de treinamento (Figura 6). Nas rodadas iniciais, a destilação do conhecimento parece retardar o progresso do modelo global, em direção ao mínimo local da função de perda, em comparação com o POC sem a KD. Ao analisar o mapa de calor de FedSeleK-Distill, a transição de preto para verde é muito mais longa do que no caso do POC. Na rodada 0, os pesos do modelo global são definidos aleatoriamente. Nas rodadas iniciais, a KD impede o modelo de fazer um progresso significativo em direção aos mínimos locais da função de perda em comparação com o POC. Portanto, o FedSeleKDistill fica atrás do POC em termos de pontuação de precisão nas primeiras rodadas. Como resultado, o POC treina o modelo global muito mais rapidamente até convergência. Isso sugere que, para conjuntos de dados locais com um nível relativamente baixo de heterogeneidade, é melhor usar o POC para obter uma convergência mais rápida.

5. Conclusão

Este artigo apresenta um novo algoritmo de treinamento dentro do paradigma de FL. O FedSeleKDistill aborda três desafios significativos no FL: o gargalo de comunicação, a heterogeneidade dos dados locais presentes nos dispositivos de clientes de FL e o número

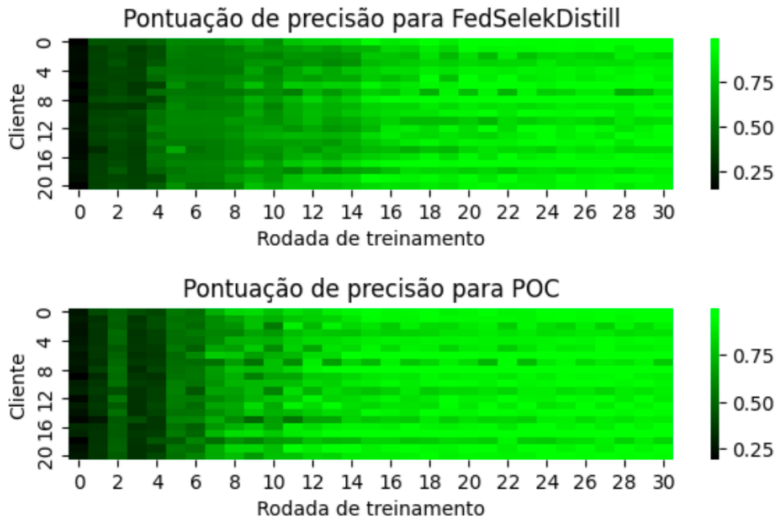

Figura 6. Pontuação de precisão do modelo global por rodada, no conjunto de dados de HAR, nos 21 clientes.

limitado de clientes disponíveis para treinamento em cada rodada. O FedSeleKDistill combina uma abordagem de seleção de cliente com a destilação de conhecimento local (KD). A primeira abordagem, baseada no *framework* POC, visa treinar o modelo global com rápida convergência. Para reduzir o risco de deriva do cliente, o FedSeleKDistill adota uma KD local no lado do cliente para mitigar o efeito negativo de dados não-IID e evitar que o modelo global esqueça o conhecimento adquirido em rodadas anteriores. As avaliações realizadas no conjunto de dados MNIST demonstram que a combinação de uma estratégia enviesada de seleção de cliente com destilação de conhecimento (KD) resulta em melhores resultados em termos de pontuações de precisão em comparação com a mesma estratégia enviesada de cliente sem a KD (o POC).

Além disso, o FedSeleKDistill requer menos rodadas de FL para treinar o modelo até a convergência em comparação com o FedAvg. A novidade da nossa solução proposta FedSeleKDistill reside na combinação de uma estratégia de seleção de cliente com a KD para mitigar o efeito negativo da heterogeneidade dos dados dos clientes. Até onde sabemos, nenhum trabalho recente na literatura explora a combinação dos dois métodos para treinamento eficiente em FL. Mesmo que tenhamos adotado uma abordagem básica de KD, os resultados experimentais mostram melhorias em comparação com o POC. Em estudos futuros, planejamos refinar nossa estratégia de KD local variando o valor de λ ao longo das rodadas de treinamento de FL, ou definindo um valor de λ dependendo do desempenho do modelo global nos dados locais dos clientes.

Agradecimentos

Este projeto foi apoiado pelo programa PPI Softex, Acordo de Parceria nº 126/2022, financiado pelo Ministério da Ciência, Tecnologia e Inovações com recursos da Lei nº 8.248, de 23 de outubro de 1991 [01245.013778/2020-21].

Referências

Amiri, M. M., Gunduz, D., Kulkarni, S. R., and Poor, H. V. (2020). Federated learning with quantized global model updates. *arXiv preprint arXiv:2006.10672*.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. (2018). signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR.

Bucila, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression.

Cho, Y. J., Wang, J., and Joshi, G. (2020). Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*.

de Souza, A. M., Bittencourt, L. F., Cerqueira, E., Loureiro, A. A., and Villas, L. A. (2023). Dispositivos, eu escolho vocês: Seleção de clientes adaptativa para comunicação eficiente em aprendizado federado. In *Anais do XLI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, 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*, page 103462.

He, Y., Chen, Y., Yang, X., Yu, H., Huang, Y.-H., and Gu, Y. (2022a). Learning critically: Selective self-distillation in federated learning on non-iid data. *IEEE Transactions on Big Data*, PP:1–12.

He, Y., Chen, Y., Yang, X., Zhang, Y., and Zeng, B. (2022b). Class-wise adaptive self distillation for federated learning on non-iid data (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12967–12968.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network.

Kaggle. Human activity recognition with smartphones.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, G., Jeong, M., Shin, Y., Bae, S., and Yun, S.-Y. (2022). Preservation of the global knowledge by not-true distillation in federated learning. *Advances in Neural Information Processing Systems*, 35:38461–38474.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2018). Federated optimization in heterogeneous networks. *CoRR*, abs/1812.06127.

Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y.-C., Yang, Q., Niyato, D., and Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063.

Lin, T., Kong, L., Stich, S. U., and Jaggi, M. (2020). Ensemble distillation for robust model fusion in federated learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2351–2363. Curran Associates, Inc.

Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, W. J. (2017). Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*.

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.

Mohamed, A. H., Assumpçáo, N. R., Astudillo, C. A., de Souza, A. M., Bittencourt, L. F., and Villas, L. A. (2023a). Compressed client selection for efficient communication in federated learning. In *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, pages 508–516. IEEE.

Mohamed, A. H., de Souza, A. M., da Costa, J. B. D., Villas, L., and dos Reis, J. C. (2023b). Ccsf: Clustered client selection framework for federated learning in non-iid data. In *Proceedings of the 16th IEEE/ACM Utility and Cloud Computing Conference (UCC)*, UCC '23, New York, NY, USA. Association for Computing Machinery.

Mora, A., Tenison, I., Bellavista, P., and Rish, I. (2022). Knowledge distillation for federated learning: a practical guide. *arXiv preprint arXiv:2211.04742*.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. (2021). A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640.

Nguyen, H. T., Sehwag, V., Hosseinalipour, S., Brinton, C. G., Chiang, M., and Poor, H. V. (2020). Fast-convergent federated learning. *IEEE Journal on Selected Areas in Communications*, 39(1):201–218.

Rothchild, D., Panda, A., Ullah, E., Ivkin, N., Stoica, I., Braverman, V., Gonzalez, J., and Arora, R. (2020). Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning*, pages 8253–8265. PMLR.

Sattler, F., Wiedemann, S., Müller, K.-R., and Samek, W. (2020). Robust and communication-efficient federated learning from non-i.i.d. data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413.

Shahid, O., Pouriyeh, S., Parizi, R. M., Sheng, Q. Z., Srivastava, G., and Zhao, L. (2021). Communication efficiency in federated learning: Achievements and challenges. *arXiv preprint arXiv:2107.10996*.

Ström, N. (2015). Scalable distributed dnn training using commodity gpu cloud computing.

Wang, H., Kaplan, Z., Niu, D., and Li, B. (2020). Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1698–1707. IEEE.

Yao, D., Pan, W., Dai, Y., Wan, Y., Ding, X., Jin, H., Xu, Z., and Sun, L. (2021). Local-global knowledge distillation in heterogeneous federated learning with non-iid data. *arXiv preprint arXiv:2107.00051*.

Zhang, L., Shen, L., Ding, L., Tao, D., and Duan, L.-Y. (2022). Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10164–10173.