طراحی زمان‌بندی و تجمیع برای یادگیری فدرال غیرهمزمان در شبکه‌های بی‌سیم

چونگ-هسوان هو، ژنگ چن، و اریک جی. لارسون

چکیده

یادگیری فدرال (FL) یک چارچوب همکاری در یادگیری ماشین (ML) است که آموزش درون‌دستگاهی را با تجمیع مبتنی بر سرور ترکیب می‌کند تا یک مدل مشترک ML را میان عوامل توزیع‌شده آموزش دهد. در این پژوهش، طراحی FL غیرهمزمان با تجمیع دوره‌ای را برای مقابله با مشکل دستگاه‌های کند (straggler) در سیستم‌های FL پیشنهاد می‌کنیم. با در نظر گرفتن محدودیت منابع ارتباطی بی‌سیم، اثر سیاست‌های زمان‌بندی مختلف و طراحی‌های تجمیع بر عملکرد همگرایی بررسی می‌شود. با تمرکز بر کاهش بایاس و واریانس به‌روزرسانی‌های مدل تجمیع‌شده، سیاست زمان‌بندی‌ای را پیشنهاد می‌کنیم که کیفیت کانال و نمایندگی داده‌های آموزشی دستگاه‌های کاربر را به‌طور مشترک در نظر می‌گیرد. اثربخشی سیاست زمان‌بندی پیشنهادی که مبتنی بر اهمیت داده و آگاهی از کانال است، در مقایسه با روش‌های پیشرفته موجود برای FL همزمان، از طریق شبیه‌سازی‌ها تأیید شده است. همچنین نشان می‌دهیم که طراحی وزن‌دهی "آگاه از سن" در تجمیع می‌تواند عملکرد یادگیری را در تنظیمات FL غیرهمزمان به‌طور چشمگیری بهبود بخشد.

**واژگان کلیدی:** یادگیری فدرال، آموزش غیرهمزمان، شبکه‌های بی‌سیم، زمان‌بندی، تجمیع

I. مقدمه

آموزش مدل‌های یادگیری ماشین (ML) معمولاً نیازمند حجم عظیمی از داده‌هاست. امروزه، افزایش روزافزون تعداد دستگاه‌های کاربر متصل، توسعه الگوریتم‌های ML را با فراهم‌سازی مجموعه‌های داده بزرگ برای آموزش مدل تسهیل کرده است. با توجه به اهمیت روزافزون حفظ حریم خصوصی، استفاده از داده‌های خصوصی دستگاه‌های کاربر برای آموزش مدل‌های ML با چالش‌هایی مواجه است. از این‌رو، یادگیری فدرال (FL) با پردازش اطلاعات درون‌دستگاهی به‌عنوان راهکاری برای حفظ حریم خصوصی داده‌ها پیشنهاد شده است. FL یک چارچوب همکاری در یادگیری ماشین است که در آن چندین دستگاه در آموزش یک مدل جهانی مشترک بر اساس داده‌های محلی خود مشارکت می‌کنند. برخلاف معماری متمرکز ML که در آن کل داده‌های آموزشی باید به‌صورت مرکزی ذخیره شوند، در سیستم FL تنها پارامترهای مدل میان دستگاه‌های کاربر و سرور پارامترها به اشتراک گذاشته می‌شود. به‌دلیل ناهمگونی دستگاه‌های مشارکت‌کننده، داده‌های آموزشی محلی ممکن است نامتوازن و غیرمستقل و غیرهم‌توزیع (non-i.i.d.) باشند، که FL را از چارچوب‌های بهینه‌سازی توزیع‌شده متعارف که در آن داده‌ها همگن و به‌طور یکنواخت توزیع شده‌اند، متمایز می‌سازد.

الگوریتم FedAvg یکی از الگوریتم‌های پایه و شناخته‌شده در FL است که شامل فرآیند تکراری انتشار مدل، آموزش محلی، و تجمیع مدل می‌باشد. در هر تکرار، فرآیند تجمیع مدل تنها زمانی آغاز می‌شود که همه دستگاه‌ها آموزش محلی را به پایان رسانده باشند. بنابراین، مدت زمان هر تکرار به‌شدت تحت تأثیر کندترین دستگاه قرار دارد. این پدیده که در روش‌های FL همزمان رایج است، به‌عنوان مشکل "straggler" شناخته می‌شود. یکی از راهکارهای مقابله با این مشکل، تغییر فرآیند همزمان به غیرهمزمان است؛ یعنی سرور نیازی ندارد تا پایان آموزش همه دستگاه‌ها صبر کند تا تجمیع را انجام دهد. در ادبیات، چنین چارچوب FL غیرهمزمانی در بسیاری از تنظیمات یادگیری عمیق مورد استفاده قرار گرفته است. با این حال، FL کاملاً غیرهمزمان با به‌روزرسانی‌های متوالی می‌تواند منجر به هزینه‌های ارتباطی بالا به‌دلیل تبادل مکرر مدل شود. از این‌رو، ما چارچوب FL غیرهمزمان با تجمیع دوره‌ای را پیشنهاد می‌کنیم که اثر straggler را حذف کرده و از به‌روزرسانی‌های بیش‌ازحد مدل و تبادل اطلاعات میان سرور و دستگاه‌ها جلوگیری می‌کند. در مقایسه با سایر کارهای موجود در زمینه FL با به‌روزرسانی‌های غیرهمزمان، طراحی پیشنهادی ما پیاده‌سازی آسانی دارد و تنها به مقدار کمی اطلاعات جانبی نیاز دارد.

محدودیت منابع ارتباطی یکی دیگر از مسائل مهم در سیستم‌های FL بی‌سیم است. از آنجا که تبادل مدل از طریق کانال‌های بی‌سیم انجام می‌شود، عملکرد سیستم (هزینه‌های ارتباطی و تأخیر) به‌طور طبیعی تحت تأثیر محدودیت منابع فرکانسی/زمانی قرار می‌گیرد، به‌ویژه زمانی که تعداد دستگاه‌های مشارکت‌کننده زیاد باشد. یکی از راه‌حل‌های کاهش بار ارتباطی این است که تنها بخشی از دستگاه‌ها برای ارسال به‌روزرسانی‌های محلی خود جهت تجمیع مدل انتخاب شوند. سپس، بسته به منابع ارتباطی تخصیص‌یافته و کیفیت لینک بی‌سیم، هر دستگاه به‌روزرسانی‌های مدل خود را به‌گونه‌ای فشرده‌سازی می‌کند که بتوان آن‌ها را با منابع تخصیص‌یافته به‌طور قابل‌اعتماد منتقل کرد. زمان‌بندی دستگاه‌ها و تخصیص منابع ارتباطی برای دستیابی به FL کارآمد از نظر ارتباطی در شبکه‌های بی‌سیم حیاتی هستند. به‌طور شهودی، دستگاه‌هایی که تأثیر بیشتری بر عملکرد یادگیری دارند باید در اولویت زمان‌بندی قرار گیرند. این هدف ارتباطی مبتنی بر یادگیری در تضاد با طراحی‌های سنتی مبتنی بر نرخ در شبکه‌های سلولی است، که هدف آن دستیابی به بازده طیفی یا توان عملیاتی بیشتر است.

II. مدل سیستم

ما یک سیستم یادگیری فدرال (FL) را در نظر می‌گیریم که در آن N دستگاه در آموزش یک مدل یادگیری جهانی مشترک مشارکت دارند. این مدل با بردار پارامتر d-بعدی θ ∈ ℝᵈ نمایش داده می‌شود. مجموعه دستگاه‌ها را با N = {1, ..., N} نشان می‌دهیم. هر دستگاه k ∈ N دارای مجموعه‌ای از داده‌های آموزشی محلی Sₖ با اندازه Sₖ است. کل مجموعه داده‌ها در سیستم با S = ⋃ₖ∈N Sₖ نمایش داده می‌شود که اندازه آن S است و Sᵢ ∩ Sⱼ = ∅ برای هر i ≠ j برقرار است. هدف سیستم یافتن بردار پارامتر بهینه θ\* است که تابع زیان تجربی زیر را کمینه کند:

F(θ) = \frac{1}{|S|} \sum\_{x ∈ S} ℓ(θ, x)

که در آن ℓ(θ, x) تابع زیان نمونه‌ای است که بر اساس نمونه داده x محاسبه می‌شود. به‌طور مشابه، تابع زیان محلی برای دستگاه k به صورت زیر تعریف می‌شود:

Fₖ(θ) = \frac{1}{|Sₖ|} \sum\_{x ∈ Sₖ} ℓ(θ, x)

بنابراین، می‌توان نوشت:

F(θ) = \sum\_{k ∈ N} \frac{|Sₖ|}{|S|} Fₖ(θ)

برای توصیف ناهمگونی توزیع داده‌ها در N، معیاری به صورت زیر تعریف می‌کنیم:

Γ = F^\* - \sum\_{k ∈ N} \frac{|Sₖ|}{|S|} Fₖ^\*

که در آن F^\* = min\_θ F(θ) و Fₖ^\* = min\_θ Fₖ(θ) هستند. در صورتی که داده‌ها به‌صورت i.i.d. توزیع شده باشند، Γ با افزایش S به صفر میل می‌کند، در حالی که در حالت non-i.i.d. انتظار داریم Γ ≠ 0 باشد که سطح ناهمگونی داده‌ها میان دستگاه‌ها را نشان می‌دهد.

A. الگوریتم FedAvg با آموزش و تجمیع همزمان

FedAvg یکی از الگوریتم‌های شناخته‌شده در یادگیری فدرال است که با رویه‌ای همزمان شامل آموزش محلی و تجمیع جهانی عمل می‌کند. کل فرآیند آموزش به چندین تکرار جهانی (دورهای ارتباطی) تقسیم می‌شود، که در هر تکرار، سرور به‌روزرسانی‌های مدل را از دستگاه‌های مشارکت‌کننده دریافت کرده و آن‌ها را تجمیع می‌کند. در اینجا نسخه‌ای اصلاح‌شده از FedAvg را در نظر می‌گیریم که در آن یک مرحله اضافی زمان‌بندی دستگاه‌ها پس از آموزش محلی اضافه شده است، همان‌طور که در شکل ۱ نشان داده شده است. هدف اصلی این مرحله کاهش هزینه‌های ارتباطی و تأخیر، به‌ویژه در شبکه‌های بی‌سیم با منابع ارتباطی محدود است.

در تکرار جهانی t، مراحل زیر اجرا می‌شوند:

1. سرور مدل جهانی فعلی θ(t) را به مجموعه دستگاه‌ها N ارسال می‌کند.
2. هر دستگاه k ∈ N به مدت E مرحله الگوریتم گرادیان تصادفی (SGD) را اجرا می‌کند. قانون به‌روزرسانی به صورت زیر است:

θₖ(t, T+1) = θₖ(t, T) - α(t, T) ∇Fₖ(θₖ(t, T); Bₖ(t, T))

که در آن T = 0, ..., E-1 شاخص مرحله محلی است، θₖ(t, 0) = θ(t)، α(t, T) نرخ یادگیری، و ∇Fₖ گرادیان محاسبه‌شده بر اساس مینی‌بچ تصادفی Bₖ(t, T) ⊆ Sₖ است. پس از پایان آموزش محلی، دستگاه k به‌روزرسانی مدل خود را به صورت زیر محاسبه می‌کند:

uₖ(t) = θₖ(t, E) - θₖ(t, 0)

1. پس از آموزش محلی، زیرمجموعه‌ای از دستگاه‌ها Π(t) ⊆ N برای ارسال به‌روزرسانی‌های خود به سرور زمان‌بندی می‌شوند.
2. پس از دریافت به‌روزرسانی‌ها از دستگاه‌های زمان‌بندی‌شده، سرور مدل جهانی را به‌صورت زیر به‌روزرسانی می‌کند:

θ(t+1) = θ(t) + \sum\_{k ∈ Π(t)} \frac{|Sₖ|}{|S|} uₖ(t)

این فرآیند تکراری تا رسیدن به همگرایی ادامه می‌یابد.

B. یادگیری فدرال غیرهمزمان با تجمیع دوره‌ای

برای رفع مشکل دستگاه‌های کند در FL همزمان بدون ایجاد بار ارتباطی بیش‌ازحد، چارچوب FL غیرهمزمان با تجمیع دوره‌ای را پیشنهاد می‌کنیم. ایده کلی این است که آموزش در دستگاه‌ها به‌صورت غیرهمزمان انجام شود، در حالی که سرور به‌طور دوره‌ای به‌روزرسانی‌ها را از دستگاه‌هایی که آموزش خود را به پایان رسانده‌اند جمع‌آوری می‌کند، و سایر دستگاه‌ها بدون وقفه به آموزش ادامه می‌دهند.

شکل ۲ تفاوت مفهومی میان FL همزمان، FL کاملاً غیرهمزمان، و طرح پیشنهادی ما را نشان می‌دهد. از آنجا که دستگاه‌ها توان پردازشی متفاوتی دارند، هرگاه دستگاهی آموزش محلی خود را به پایان رساند، سیگنالی به سرور ارسال می‌کند تا آمادگی خود را برای ارسال به‌روزرسانی اعلام کند. در هر بازه زمانی ثابت T، سرور زیرمجموعه‌ای از دستگاه‌های آماده را زمان‌بندی می‌کند. مجموعه دستگاه‌های آماده در تکرار t را با K(t) نشان می‌دهیم. مجموعه دستگاه‌های زمان‌بندی‌شده Π(t) ⊆ K(t) است، با شرط Π(t) = min{R, K(t)}، که R حداکثر تعداد دستگاه‌های قابل زمان‌بندی است.

برای جلوگیری از عدم تعادل مدل ناشی از آموزش غیرهمزمان، یک جمله منظم‌سازی به تابع زیان اضافه می‌کنیم. تابع زیان منظم‌شده محلی برای دستگاه k در تکرار t و مرحله T به صورت زیر تعریف می‌شود:

Hₖ(t, T) = Fₖ(θₖ(t, T)) + \frac{λ}{2} ||θₖ(t, T) - θₖ(t, 0)||²

که در آن λ > 0 ضریب منظم‌سازی است. به‌روزرسانی محلی برای دستگاه k ∈ Π(t) به صورت زیر انجام می‌شود:

θₖ(t, T+1) = θₖ(t, T) - α(t, T) ∇Hₖ(t, T)

با T = 0, ..., E-1 و گرادیان ∇Hₖ(t, T) که بر اساس مینی‌بچ تصادفی Bₖ(t, T) ⊆ Sₖ محاسبه می‌شود. به‌دلیل تنظیم غیرهمزمان، مدل اولیه در مرحله اول آموزش محلی به صورت زیر تعریف می‌شود:

θₖ(t, 0) = θ(Sₖ(t))

که در آن Sₖ(t) آخرین تکرار جهانی است که دستگاه k مدل به‌روزرسانی‌شده را دریافت کرده است. پس از دریافت به‌روزرسانی‌ها از دستگاه‌های Π(t)، سرور مدل را به‌صورت زیر تجمیع می‌کند:

θ(t+1) = \sum\_{k ∈ Π(t)} wₖ(t) [θₖ(t, 0) + uₖ(t)]

که در آن uₖ(t) همان‌طور که پیش‌تر تعریف شد، و wₖ(t) ضریب وزن‌دهی است، با شرط ∑ₖ∈Π(t) wₖ(t) = 1. مدل به‌روزرسانی‌شده سپس به تمام دستگاه‌های K(t) ارسال می‌شود تا آموزش محلی ادامه یابد.

با الهام از مفهوم "سن اطلاعات" (Age of Information)، سن به‌روزرسانی محلی (ALU) را به صورت زیر تعریف می‌کنیم:

aₖ(t) = t - Sₖ(t)

که نشان‌دهنده مدت زمان سپری‌شده از آخرین دریافت مدل به‌روزرسانی‌شده توسط دستگاه k است. وزن wₖ(t) می‌تواند بر اساس اندازه داده‌های آموزشی Sₖ، یا ALU aₖ(t)، یا ترکیبی از هر دو تعیین شود.

C. مدل فیزیکی لایه ارتباطی (PHY)

در فرآیند یادگیری فدرال، ارتباطات در دو مرحله انجام می‌شود: ارسال مدل جهانی از سرور به دستگاه‌ها (ارسال پایین‌رو) و ارسال به‌روزرسانی‌های مدل از دستگاه‌ها به سرور (ارسال بالا‌رو). به‌طور کلی، توان ارسال سرور بسیار بیشتر از دستگاه‌هاست و ارسال پایین‌رو به‌دلیل ماهیت کانال پخش، نیاز به تقسیم منابع ارتباطی ندارد. بنابراین، فرض می‌کنیم در ارسال پایین‌رو خطایی در فشرده‌سازی وجود ندارد. با این حال، ارسال بالا‌رو با محدودیت منابع ارتباطی مواجه است، که طراحی زمان‌بندی و تخصیص منابع را بسیار مهم می‌سازد.

ما فرض می‌کنیم که در هر تکرار جهانی t، به‌روزرسانی‌های مدل از همه دستگاه‌ها از طریق کانال محو بلوکی با بلوک هم‌پایگی n نماد منتقل می‌شوند. تخصیص منابع به‌صورت متعامد میان دستگاه‌ها انجام می‌شود، یعنی به هر دستگاه k تعداد nₖ(t) نماد اختصاص داده می‌شود، به‌طوری‌که مجموع آن‌ها برابر با n است. بنابراین، انتقال‌ها از دستگاه‌های مختلف بدون تداخل هستند. پارامترهای Bₖ و hₖ(t) به‌ترتیب محو بزرگ‌مقیاس و کوچک‌مقیاس کانال از دستگاه k به سرور را نشان می‌دهند. با فرض توان ارسال Pₖ(t) ≤ P\_max و نویز افزوده با توزیع نرمال مختلط CN(0, σ²)، ظرفیت کانال برای دستگاه k به‌صورت زیر است:

Cₖ(t) = \log\_2 \left(1 + \frac{Pₖ(t) Bₖ |hₖ(t)|²}{σ²} \right)

فرض می‌کنیم دستگاه‌ها از طرح‌های فشرده‌سازی داده و کدگذاری کانال مناسب استفاده می‌کنند تا انتقال به‌روزرسانی‌های مدل بدون خطا انجام شود. در نتیجه، سرور می‌تواند تا nₖ(t)·Cₖ(t) بیت از دستگاه k دریافت کند. تخصیص نمادها و طرح فشرده‌سازی داده‌ها به‌شرح زیر است:

1. تخصیص منابع ارتباطی: برای دستیابی به سطح یکسانی از خطای فشرده‌سازی در به‌روزرسانی‌های مدل از هر دستگاه، نمادها به‌گونه‌ای تخصیص می‌یابند که:

n₁(t)·C₁(t) = ... = n\_{|Π(t)|}(t)·C\_{|Π(t)|}(t)

که به این معناست که دستگاه‌هایی با کیفیت کانال بهتر، نمادهای کمتری دریافت می‌کنند.

1. تنک‌سازی و کم‌دقت‌سازی: هر دستگاه باید uₖ(t) را مطابق با بودجه nₖ(t)·Cₖ(t) بیت فشرده کند. از تنک‌سازی تصادفی و سپس کم‌دقت‌سازی تصادفی با سطح v استفاده می‌کنیم. به‌طور خاص، بردار فشرده‌شده ūₖ(t) ∈ ℝᵈ به‌صورت زیر تعریف می‌شود:

\hat{u}\_{k,i}(t) =

\begin{cases}

Q\_v(u\_{k,i}(t)) & \text{اگر } i ∈ Vₖ(t) \\

0 & \text{در غیر این صورت}

\end{cases}

که در آن Vₖ(t) مجموعه‌ای از عناصر انتخاب‌شده با اندازه rₖ(t) است، و Q\_v(.) کم‌دقت‌ساز تصادفی سطح v است. این کم‌دقت‌ساز بدون بایاس است و واریانس آن محدود است. مقدار rₖ(t) به‌گونه‌ای انتخاب می‌شود که قید انتقال زیر را ارضا کند:

\log\_2 \binom{d}{rₖ(t)} + 32 + rₖ(t)·(\log\_2(v+1)+1) ≤ nₖ(t)·Cₖ(t)

که در آن:

* اولین جمله برای ارسال اندیس‌هاست،
* ۳۲ بیت برای ارسال نُرم بردار است،
* و جمله سوم برای ارسال عناصر غیرصفر با کم‌دقت‌سازی است.

D. انگیزه طراحی زمان‌بندی و تجمیع

در چارچوب FL غیرهمزمان پیشنهادی، دو پرسش کلیدی طراحی مطرح می‌شود:

1. با توجه به محدودیت منابع ارتباطی، چگونه باید زیرمجموعه‌ای از K(t) را برای تجمیع مدل زمان‌بندی کرد، با در نظر گرفتن توزیع داده‌های آموزشی ناهمگن و کیفیت لینک بی‌سیم؟
2. دستگاه‌های مختلف ممکن است ALU متفاوتی داشته باشند — برخی به‌روزرسانی‌های تازه‌تر و برخی قدیمی‌تر. چگونه باید سیاست وزن‌دهی مناسبی طراحی کرد که تازگی به‌روزرسانی‌ها را در نظر بگیرد؟

هدف، کمینه‌سازی تابع زیان تجربی جهانی F(θ) است. فرض می‌کنیم F(θ) قویاً محدب و L-نرم باشد. در این صورت، در تکرار (t+1) داریم:

E[F(θ(t+1))] - F^\* ≤ \frac{L}{2} E[||θ(t+1) - θ^\*||²]

که در آن E[.] امید ریاضی نسبت به تصادفی بودن مراحل گذشته است. جمله سمت راست می‌تواند به‌عنوان واریانس مدل تجمیع‌شده تفسیر شود. عوامل مؤثر بر آن عبارت‌اند از:

1. توزیع داده‌های آموزشی: در حالت i.i.d.، انتظار داریم Γ → 0، و بنابراین واریانس مدل تجمیع‌شده کوچک باشد. اما در حالت non-i.i.d.، تفاوت میان مدل‌های محلی زیاد است و واریانس افزایش می‌یابد.
2. فشرده‌سازی داده‌ها: فشرده‌سازی در کانال‌های محدود باعث دریافت اطلاعات نویزی توسط سرور می‌شود، که واریانس مدل را افزایش می‌دهد.
3. به‌روزرسانی‌های غیرهمزمان: تفاوت در ALU میان دستگاه‌ها منبع اضافی واریانس است.

از دیدگاه زمان‌بندی دستگاه‌ها، می‌توان با انتخاب دستگاه‌هایی که نمایندگی همگنی از کل داده‌ها دارند و کیفیت کانال بهتری دارند، واریانس را کاهش داد. از دیدگاه تجمیع مدل، می‌توان با طراحی وزن‌دهی آگاه از سن، اثرات منفی به‌روزرسانی‌های غیرهمزمان را کاهش داد.

III. طراحی زمان‌بندی و تجمیع برای یادگیری فدرال غیرهمزمان

ما سیاست زمان‌بندی‌ای را پیشنهاد می‌کنیم که هدف آن کاهش شکاف بهینگی است؛ این سیاست شرایط کانال و توزیع داده‌های آموزشی دستگاه‌های آماده به‌روزرسانی را در نظر می‌گیرد. سپس، وزن‌های تجمیع به‌گونه‌ای تنظیم می‌شوند که اثرات منفی ناشی از آموزش غیرهمزمان کاهش یابد.

A. زمان‌بندی مبتنی بر اهمیت داده و آگاهی از کانال

برای توضیح ایده، یک مسئله طبقه‌بندی با داده‌های برچسب‌خورده را در نظر می‌گیریم. مجموعه داده‌های آموزشی به‌صورت S = {(x, y) y ∈ ℒ} نمایش داده می‌شود، که ℒ = {y₁, ..., y\_w} مجموعه‌ای متناهی از برچسب‌هاست و ℒ = w. برای هر دستگاه k، بردار bₖ = [b₁, ..., b\_w] توزیع برچسب‌ها در Sₖ را نشان می‌دهد، که در آن bⱼ تعداد نمونه‌های دارای برچسب yⱼ است و مجموع bⱼ برابر با Sₖ است.

برای ساخت توزیع داده‌ای همگن، می‌توان مجموعه Π(t) را به‌گونه‌ای انتخاب کرد که واریانس برچسب‌ها کمینه شود:

Q(Π(t)) = \sum\_{j=1}^{w} \left( \sum\_{k ∈ Π(t)} b\_{k,j} - b\_j \right)^2

که در آن b\_j میانگین توزیع برچسب‌ها در کل دستگاه‌هاست. علاوه بر این، زمان‌بندی دستگاه‌هایی با کیفیت کانال بهتر منجر به کاهش خطای فشرده‌سازی می‌شود. با ترکیب این دو معیار، ابتدا زیرمجموعه Π'(t) ⊆ K(t) شامل دستگاه‌هایی با بالاترین ظرفیت کانال Cₖ(t) انتخاب می‌شود (حداکثر ۰٫۵N یا K(t)). سپس، مجموعه Π(t) به‌گونه‌ای انتخاب می‌شود که Q(Π(t)) کمینه شود، یعنی:

Π(t) = \arg\min\_{Π ⊆ Π'(t)} Q(Π)

B. تجمیع مدل آگاه از سن

برای مقابله با اثرات منفی تجمیع غیرهمزمان، وزن‌دهی به‌روزرسانی‌ها نه‌تنها بر اساس اندازه داده‌ها، بلکه بر اساس سن به‌روزرسانی (ALU) انجام می‌شود. طراحی وزن‌دهی آگاه از سن به‌صورت زیر است:

wₖ(t) = \frac{|Sₖ| · aₖ(t)^γ}{\sum\_{j ∈ Π(t)} |Sⱼ| · aⱼ(t)^γ}, \quad ∀k ∈ Π(t)

که در آن γ یک عدد حقیقی است و انتخاب مقدار آن به سه حالت تقسیم می‌شود:

* γ > 1: سیستم به‌روزرسانی‌های قدیمی‌تر را ترجیح می‌دهد.
* γ < 1: سیستم به‌روزرسانی‌های تازه‌تر را ترجیح می‌دهد.
* γ = 1: معادل طراحی پایه‌ای وزن‌دهی بر اساس اندازه داده‌هاست.

ترجیح دادن به‌روزرسانی‌های قدیمی‌تر می‌تواند مشارکت دستگاه‌های ضعیف‌تر را متعادل کند، به‌ویژه زمانی که آن‌ها داده‌های منحصربه‌فردی دارند. اما این روش ممکن است باعث اعمال به‌روزرسانی‌های منسوخ بر مدل فعلی شود. در مقابل، ترجیح دادن به‌روزرسانی‌های تازه‌تر باعث همگرایی سریع‌تر و روان‌تر مدل می‌شود، هرچند ممکن است مدل نهایی به داده‌های دستگاه‌های قدرتمندتر متمایل شود. از آنجا که سیاست زمان‌بندی پیشنهادی ما همگنی داده‌ها را بهبود می‌بخشد، استراتژی ترجیح مدل‌های تازه‌تر منطقی‌تر به‌نظر می‌رسد.

در مجموع، طراحی زمان‌بندی مبتنی بر اهمیت داده و کیفیت کانال همراه با سیاست تجمیع ترجیح‌دهنده به‌روزرسانی‌های تازه، منجر به عملکرد همگرایی سریع و روان می‌شود. برای اثبات اثربخشی این طراحی، در بخش IV تحلیل همگرایی و در بخش V نتایج شبیه‌سازی ارائه شده‌اند.

IV. تحلیل همگرایی

برای تحلیل همگرایی سیستم پیشنهادی، ابتدا برخی نمادها و تعاریف را معرفی می‌کنیم.

تعریف ۱: (ناهمزمانی در تجمیع)

* فرض کنید M(t) تعداد نسخه‌های مختلف مدل جهانی دریافت‌شده در تکرار t باشد، یعنی تعداد عناصر منحصربه‌فرد در مجموعه {aₖ(t) k ∈ Π(t)}.
* مجموعه Mᵢ(t) ⊆ Π(t)، برای i = 1, ..., M(t)، زیرمجموعه‌ای از دستگاه‌ها با ALU یکسان است، یعنی برای هر k، j ∈ Mᵢ(t)، داریم aₖ(t) = aⱼ(t).

از آنجا که زمان‌بندی دستگاه‌ها در سیستم پیشنهادی بر اساس مجموعه آماده به‌روزرسانی K(t) انجام می‌شود، نه کل مجموعه N، برای هر زیرمجموعه M از دستگاه‌ها، معیارهای زیر برای سنجش سطح ناهمگونی داده‌ها تعریف می‌شوند:

تعریف ۲: (سطح ناهمگونی داده‌ها)

برای زیرمجموعه‌ای از دستگاه‌ها M ⊆ N و وزن‌های wₖ با مجموع ۱، داریم:

T₁(M) = F^\* - \sum\_{k ∈ M} wₖ Fₖ

T₂(M) = F(θ^\*) - \sum\_{k ∈ M} wₖ Fₖ(θ^\*)

A. فرضیات

برای تسهیل تحلیل، فرضیات زیر را در نظر می‌گیریم:

* فرض ۱ (نرمی): تابع زیان محلی Fₖ(θ) برای هر k، L-نرم است، یعنی:

||∇Fₖ(θ₁) - ∇Fₖ(θ₂)|| ≤ L ||θ₁ - θ₂||

* فرض ۲ (قویاً محدب بودن): تابع Fₖ(θ) برای هر k، μ-قویاً محدب است، یعنی:

Fₖ(θ₁) - Fₖ(θ₂) ≥ ∇Fₖ(θ₂)^T (θ₁ - θ₂) + \frac{μ}{2} ||θ₁ - θ₂||²

* فرض ۳ (محدودیت دومین لحظه): برای هر θ ∈ ℝᵈ، گرادیان تصادفی ∇Fₖ(θ; Bₖ) دارای محدودیت زیر است:

E[||∇Fₖ(θ; Bₖ)||²] ≤ C₁ + C₂ ||∇Fₖ(θ)||²

که در آن C₁ و C₂ مقادیر مثبت هستند.

نکته: در بسیاری از مقالات، فرض می‌شود گرادیان‌ها به‌طور یکنواخت محدود هستند (یعنی C₂ = ۰)، در حالی که این فرض با فرض قویاً محدب بودن در فضای نامحدود ناسازگار است. بنابراین، ما فرض می‌کنیم نویز گرادیان‌ها در امید ریاضی محدود است، که با فضای نامحدود سازگارتر است.

نکته ۱

با تجمیع دوره‌ای با دوره T، ALU هر دستگاه با زمان آموزش محدود به T\_max، در بازه ۰ تا ⌊T\_max / T⌋ قرار دارد. بنابراین:

M(t) ≤ \left\lfloor \frac{T\_{\text{max}}}{T} \right\rfloor + 1

نکته ۲

برای هر زیرمجموعه M از N، چون N محدود است، مقادیر T₁(M) و T₂(M) به‌طور یکنواخت محدود هستند:

|T₁(M)| ≤ τ₁, \quad |T₂(M)| ≤ τ₂

فرض ۴

برای تعداد عناصر انتخاب‌شده rₖ(t) در فشرده‌سازی، فرض می‌کنیم مقدار میانگین آن محدود از پایین باشد:

E[rₖ(t)] ≥ r\_{\text{min}} > 0

B. مطالعه همگرایی

در ادامه، نتیجه اصلی همگرایی را در حالت خاصی ارائه می‌کنیم که در آن تنها یک مرحله آموزش محلی در هر تکرار جهانی انجام می‌شود (یعنی E = 1).

قضیه ۱

با فرضیات ۱ تا ۳، روابط بالا، و نرخ یادگیری کاهش‌یابنده α(t, 0) = B / (t + K + 1) با B > 0، و K = max{(B r\_min - 1)/(4 L B), 1}، و مشارکت جزئی دستگاه‌ها به‌صورت Π(p) = ⋃\_{i=1}^{M(p)} Mᵢ(p) ⊆ N برای p = 1, ..., t، سیستم پیشنهادی رابطه زیر را ارضا می‌کند:

E[F(θ(t+1))] - F^\* ≤ \frac{L}{2(t + K + 1)} \left[ J(B, K) + (K + 1) ||θ(1) - θ^\*||² \right] + \frac{B A}{t + K + 1}

که در آن:

* A = 2(τ₁ + τ₂)
* C₃ = 4(τ₁ + τ₂)
* J(B, K) = [B² C₃ (τ₁ + A L C₂) + L² C₂ C₃ A] / (d (B r\_min - 1))

امید ریاضی نسبت به تصادفی بودن گرادیان‌ها، کانال‌ها، فشرده‌سازی، و زمان‌بندی دستگاه‌ها در تمام تکرارهای گذشته گرفته شده است.

نکته ۳

تأثیر چند مرحله آموزش محلی (E > 1) در تحلیل همگرایی در مقالات پیشین بررسی شده است. تمرکز تحلیل ما بر تأثیر فشرده‌سازی، ناهمگونی داده‌ها، و غیرهمزمانی درون‌تکراری است. گسترش نتایج به حالت E > 1 ممکن است با استفاده از روش‌های موجود، اما نیازمند تحلیل واگرایی مدل‌های محلی خواهد بود.

نکته ۴

بر اساس قضیه ۱، شکاف بهینگی به‌صورت زیر به مقدار ثابتی همگرا می‌شود:

\lim\_{t → ∞} E[F(θ(t+1))] - F^\* = \frac{B A}{L M}

که در آن M، A، و B همان‌طور که در بالا تعریف شده‌اند. این مقدار ثابت را می‌توان به‌عنوان حد نهایی شکاف بهینگی در نظر گرفت.

V. نتایج شبیه‌سازی

برای ارزیابی عملکرد طراحی پیشنهادی، شبیه‌سازی‌هایی در محیط‌های مختلف انجام شده است. در این شبیه‌سازی‌ها، ما یک مسئله طبقه‌بندی را با داده‌های غیرهم‌توزیع (non-i.i.d.) میان دستگاه‌ها در نظر گرفته‌ایم. پارامترهای سیستم شامل تعداد دستگاه‌ها، ظرفیت کانال‌ها، نرخ یادگیری، و سطح فشرده‌سازی تنظیم شده‌اند تا شرایط واقعی شبکه‌های بی‌سیم را بازتاب دهند.

نتایج نشان می‌دهند که:

* چارچوب FL غیرهمزمان با تجمیع دوره‌ای، در مقایسه با روش‌های همزمان، همگرایی سریع‌تری دارد و از تأخیر ناشی از دستگاه‌های کند جلوگیری می‌کند.
* سیاست زمان‌بندی مبتنی بر اهمیت داده و کیفیت کانال، منجر به کاهش واریانس مدل تجمیع‌شده و بهبود عملکرد نهایی می‌شود.
* طراحی وزن‌دهی آگاه از سن، به‌ویژه در حالتی که γ < 1 باشد (ترجیح به‌روزرسانی‌های تازه‌تر)، باعث همگرایی روان‌تر و مدل متوازن‌تر می‌شود.

در مجموع، ترکیب زمان‌بندی هوشمند و تجمیع آگاه از سن، عملکرد یادگیری را در محیط‌های غیرهمگن و با منابع ارتباطی محدود به‌طور چشمگیری بهبود می‌بخشد.

VI. جمع‌بندی

در این مقاله، یک چارچوب یادگیری فدرال غیرهمزمان با تجمیع دوره‌ای برای مقابله با مشکل دستگاه‌های کند در سیستم‌های FL پیشنهاد شد. طراحی پیشنهادی با در نظر گرفتن محدودیت‌های ارتباطی و ناهمگونی داده‌ها، شامل:

* سیاست زمان‌بندی مبتنی بر کیفیت کانال و اهمیت داده‌ها،
* طراحی وزن‌دهی آگاه از سن برای تجمیع مدل،
* تحلیل همگرایی با در نظر گرفتن فشرده‌سازی، ناهمگونی داده‌ها، و به‌روزرسانی‌های غیرهمزمان.

نتایج نظری و تجربی نشان دادند که این طراحی منجر به همگرایی سریع‌تر، کاهش واریانس مدل، و عملکرد بهتر در محیط‌های واقعی می‌شود.

VII. ملاحظات نهایی

طراحی پیشنهادی نه‌تنها برای یادگیری فدرال در شبکه‌های بی‌سیم کاربرد دارد، بلکه می‌تواند در سایر تنظیمات یادگیری توزیع‌شده با داده‌های ناهمگن و منابع محدود نیز مورد استفاده قرار گیرد. در آینده، می‌توان این چارچوب را با روش‌های پیشرفته فشرده‌سازی، تخصیص منابع پویا، و یادگیری تطبیقی ترکیب کرد تا عملکرد آن در شرایط پیچیده‌تر نیز حفظ شود.