Optimizing IoT Data Collection for Federated Learning Under Constraint of Wireless Bandwidth

竹本志恩

April 18, 2025

INIAD

目次

書誌情報

- **題名**
  - Optimizing IoT Data Collection for Federated Learning Under Constraint of Wireless Bandwidth
- **発表日**
  - 20 August 2024
- **著者**
  - Tajiri Kengo, Ryoichi Kawahara
- **論文誌名**
  - IEEE

どんな研究?

- IoT でのデータ利活用のため
- 帯域幅制約の下で実験を行い
- 提案手法による FL* の精度向上を確認

  * Federated Learning, 連合学習

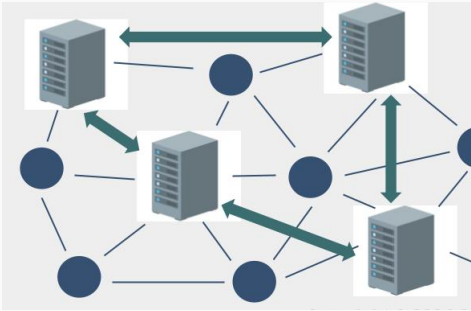
目次

背景

- IoT の普及に伴いデバイスが増加
- 従来のデータ集約・分析が困難に
  - 総データ生成量の増加
  - 利用帯域・計算コストの増加
- Federated Learning に着目
  - 分散機械学習の一手法
  - データを分散し, 上記の問題を解決

Federated Learning とは

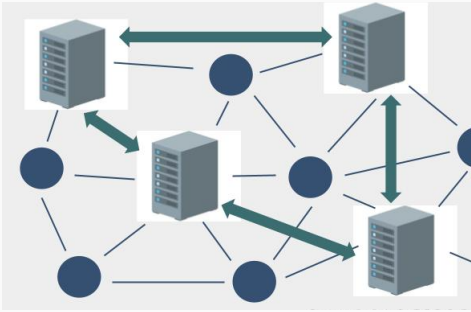
- 連合学習, FL
  - 分散的に学習する手法の一つ
  - 各サーバはデータを持ち, 学習
  - 学習結果をあるサーバで集約
  - 統合, パラメータサーバが存在
- 大まかな流れ
  - 統合サーバからモデルを配布
  - 各パラメータサーバはローカルデータで学習
  - モデルのパラメータを統合サーバで集約
  - モデルを更新し再配布
  - 一連の流れを繰り返す



FL のイメージ

課題

- IoT デバイス (ID) から基地局 (BS) へのデータ転送
- 想定されるシナリオ
  - 各 BS はサーバと ID を所持
  - BS に ID からデータを送信
  - サーバの持つデータでモデルを学習
- FL の精度
  - サーバの持つデータの量と分布が影響
  - 量と分布は制約の下決めた伝送先に依存
  - ID からどの BS に送信するかが重要
- 帯域幅制約下で FL の精度を向上
  - 制約により送信先が限定
  - その中で最適な BS を決定



再掲

目次

手法の概要

- やること
  - 帯域制約の下
  - FL の精度を最大化するような
  - 最適化問題を解く
- 解法
  - 遺伝的アルゴリズム
  - ID からどの BS にデータを送信するか決定
- 実験
  - 提案手法とナイーブを比較
  - 数値実験
  - 仮想クラスタでの学習
    - 三種類の機械学習モデルで比較

OFDMA

- 周波数を分割し割当て
- 単一の通信路で複数のやり取りが可能に
- リソースブロック (RB) という単位で管理?
  - データのやり取りに用いるそういう塊がある

問題設定

- ID: $I = \{1, 2, ..., N\}$
- BS: $B = \{1, 2, ..., M\}$
- **単位時間あたりのデータ生成量**: λ_i
- **一回の学習におけるデータ収集量**: T
- RB: $R = \{1, 2, ..., K\}$

制約条件 1-4

- ID が RB を使って BS にデータを渡すよ，というのを定義
- ID に対して BS は 0 個でもいい
  - 送らない ID の存在を許容
- 伝送時には BS と RB が同じだけ必要
- RB は ID に一つのみ割り当て
  - 送らない場合があるので不等式

$$\sum_{j\in\mathcal{B}} u_{ij} \leq 1 \quad \forall i \in \mathcal{I}, \tag{1}$$
$$\sum_{k\in\mathcal{R}} v_{ik} \leq 1 \quad \forall i \in \mathcal{I}, \tag{2}$$
$$\sum_{j\in\mathcal{B}} u_{ij} = \sum_{k\in\mathcal{R}} v_{ik} \quad \forall i \in \mathcal{I}. \tag{3}$$

$$\sum_{i\in\mathcal{I}} u_{ij} v_{ik} \leq 1 \quad \forall j \in \mathcal{B}, \forall k \in \mathcal{R}. \tag{4}$$

制約条件 5-8

- 無線接続の帯域幅を定義
- 7 が単純に ID BS 間の帯域幅
- 式 8 が基本の帯域幅制約
  - 8 は混雑を回避するための制約
  - 7 で定めた帯域幅がある伝送路を通る総データ生成量を以上なる

$$c_{ij}^{k} = B \log_2(1 + \gamma_{ij}^{k}), \tag{7}$$

$$u_{ij} c_{ij}^{k} \geq u_{ij} v_{ik} \lambda_i \quad \forall i \in \mathcal{I}, \forall j \in \mathcal{B}, \forall k \in \mathcal{R}. \tag{8}$$

制約条件 9-18

- 汎化誤差の上界を推定
  - 汎化誤差は未知のデータに対する予測の誤差
  - 上界は実際の集合より大きな数値のこと?
- 汎化誤差を最小化したい
  - 予測の精度を向上させるのが目的
  - 汎化誤差を最小化する式は 18
  - パラメータは U,VID に対して BS は 0 個でもいい

$$F(U,V) = p(c)\log p(c) - \sum_{j\in\mathcal{B}} \sigma_j \sum_{c\in\mathcal{C}} p(c)\log p_j(c) + \beta\sqrt{\frac{\log|D|}{2|D|}}, \tag{17}$$

$$\min_{U,V}(17) \qquad \text{s.t.}(1)-(4),(8) \tag{18}$$

実験1 数値シミュレーション

- 問題設定
  - ID 20,BS 3,RB 5
  - ID, BS は 1km の正方形空間にランダム配置
  - ID で生成されるデータの量: $\lambda_i = 100|1000$ Kbps
  - 学習時間の T における総量 50000 データ
  - ラベル数は 10
  - 各 ID のラベル分布 $p_i(C)$ はランダム
- 生成データ量を変えて実験
- 式 17 の β 依存性も見た
- データの量と分布を確認
- 最適化は 4,8 の制約に従う

$$\sum_{i \in \mathcal{I}} u_{ij} v_{ik} \leq 1 \quad \forall j \in \mathcal{B}, \forall k \in \mathcal{R}.$$

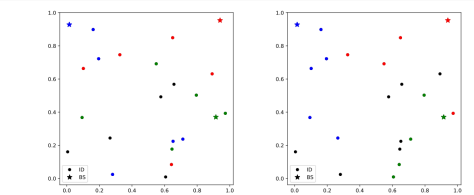
$$u_{ij} c_{ij}^k \geq u_{ij} v_{ik} \lambda_i \quad \forall i \in \mathcal{I}, \forall j \in \mathcal{B}, \forall k \in \mathcal{R}.$$

実験1 数値シミュレーション

- ナイーブ法
  - ID はランダムに配置
  - 最も近い BS に接続

実験１の結果

- fig2 の見方
  - 星が BS
  - 点が ID
  - 同じ色の BS に接続
  - 黒は未接続



(a) proposed optimization with $\bar{\lambda} = 100$ Kbps and $\beta = 10$

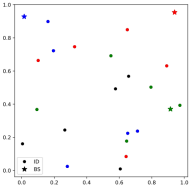
(b) proposed optimization with $\bar{\lambda} = 1000$ Kbps and $\beta = 10$

(c) naive method with $\bar{\lambda} = 100$ Kbps

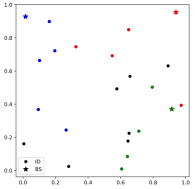
(d) naive method with $\bar{\lambda} = 1000$ Kbps

実験 1 の結果

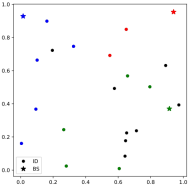
- 図 a,b の比較
  - データ生成量の多寡
  - a は生成量小
    - 制約 8 での送信先の縛りが緩い
  - b は生成量大
    - 扱うデータ量が増大
    - 制約がより厳しくなる
    - 近い BS に送信することが多い
- 提案手法とナイーブの比較
  - 提案手法は BS に接続した ID の総量が多い
  - 利用したデータの数が多い
  - KL(p(c)∣∣ pj (c)) が小さい?



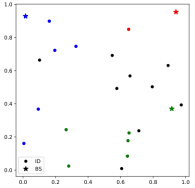
(a) proposed optimization with $\bar{\lambda} = 100$ Kbps and $\beta = 10$



(b) proposed optimization with $\bar{\lambda} = 1000$ Kbps and $\beta = 10$



(c) naive method with $\bar{\lambda} = 100$ Kbps



(d) naive method with $\bar{\lambda} = 1000$ Kbps

実験 2 深層学習モデル 問題設定

- データセット
  - cifar10
  - 画像の 10 クラス分類
  - 全部で 5 万のデータ
- ID のデータ分布
  - シミュレーションで決まった λ i と pi (c) に従う
  - ランダムに選択?
- 利用したモデル
  - VGG19
  - DenseNet121
  - ResNet152

実験 2 深層学習モデル 問題設定

- FL を Flower で実装
- FedAVG で集約
- FL の設定
  - 5 つのローカルトレーニング
  - 多分五回
  - 50 エポック
  - 一回で 50 回?
  - 50 回になるよう繰り返した?

実験 2 結果

- データセット
- 精度の差が出た理由
  - 100 の時
    - 2 ポイントほど差があった
    - 制約が緩いため
    - 最良のデータ転送が行えた
    - 量と分布が最適になる転送?
  - データ生成量 1000 の時
    - 精度はほぼ同じ
    - データが増え, 制約が厳しく
- β も KL も最適化で重要
- モデルの構造に依存する?

TABLE II: Results of simulation and actual models accuracy. ac

	proposed optimization			
$\bar{\lambda}$	100	1000	1000	1000
β	10	10	1	100
$\sum_{i\in\mathcal{I},j\in\mathcal{B}} u_{ij}$	15	13	12	14
$KL(p(c)\|\|p_j(c))$	7.724×10^{-3}	1.127×10^{-2}	6.643×10^{-3}	4.260×1
$\|D\|$	39015	33189	29565	33585
VGG19 acc.	0.8072±0.0006	0.7955±0.0085	0.7813±0.0053	0.7966±0.
DenseNet121 acc.	0.8301±0.0015	0.8151±0.0036	0.8051±0.0006	0.8206±0.
ResNet152 acc.	0.7423±0.0033	0.7244±0.0064	0.7098±0.0021	0.7232±0.

実験2 結果

TABLE II: Results of simulation and actual models accuracy. acc. means accuracy.

	proposed optimization				naive method	
$\bar{\lambda}$	100	1000	1000	1000	100	1000
β	10	10	1	100		
$\sum_{i\in\mathcal{I},j\in\mathcal{B}} u_{ij}$	15	13	12	14	12	11
$KL(p(c)\|\|p_j(c))$	7.724×10^{-3}	1.127×10^{-2}	6.643×10^{-3}	4.260×10^{-2}	5.463×10^{-2}	5.339×10^{-2}
$\|D\|$	39015	33189	29565	33585	30150	29369
VGG19 acc.	0.8072±0.0006	0.7955±0.0085	0.7813±0.0053	0.7966±0.0033	0.7857±0.0012	0.7793±0.0066
DenseNet121 acc.	0.8301±0.0015	0.8151±0.0036	0.8051±0.0006	0.8206±0.0021	0.8099±0.0044	0.8139±0.0069
ResNet152 acc.	0.7423±0.0033	0.7244±0.0064	0.7098±0.0021	0.7232±0.0027	0.7164±0.0050	0.7181±0.0052

目次

得られた結果, 知見

- 提案手法はナイーブと比べて精度を改善
  - いずれの実験でも改善された
    - 実験 2 は平均値で観察
    - 詳細な設定はまだ見ていない
  - データ転送の最適化で FL の精度が向上
  - ID から BS にデータを送信し FL を行うシナリオにおいて
    - 提案手法は制約条件の下, データ量と分布が最適になるようデータを送信
    - データ転送の観点から精度の向上が実現

目次

分かっていない/気になる点

- 用語など
  - 遺伝的アルゴリズム
  - 最適化の式
- 精度がナイーブより高いことの意味
  - 制約を守った上で精度が高いのが偉い?
  - 帯域幅の制約を守る FL が先??
  - 守った上で精度が出せたのが偉いのか, 精度が出せるのは当たり前なのか
- 数値シミュレーション
  - 最適化手法は式 17 を小さくするために多数接続
  - BS に接続する ID の数が増えると良いらしい
  - β 依存性の議論

分かっていない/気になる点

- より複雑なモデルでの検証
- ID,BS 間の具体的な通信
- FL について, バッテリー制約も同時に満たせないか
- モデル転送も考慮できないか