Data Valuation for Machine Learning and Federated Learning

Student Name: CHEN Jiaqing

Supervisor: Dr. WANG Cong

Content

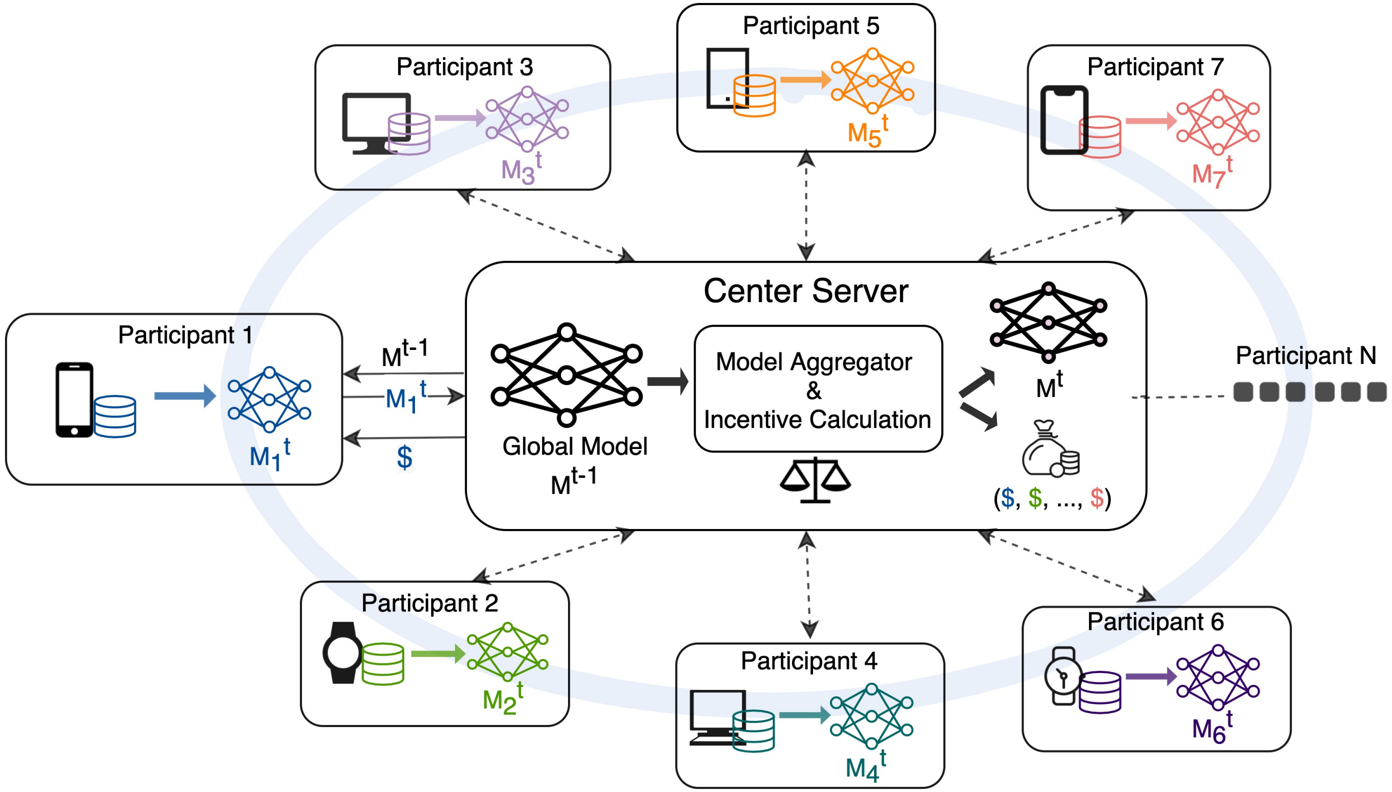
1.1 Federated Learning (FL): Why and What?

- The popularity of machine learning (ML)
- Data is important, but dispersed in different places
- Directly access → incur privacy issues

Federated Learning:

A solution to

- gather isolated data
- perform collaborative model training
- mitigates the privacy risks from being transmitted between local users and the center



1.2 Incentive Scheme

Motivation: Encourage long-term participations

Objective: Rewards should be proportional to the local user's contribution

- High quality dataset → high reward
- Random data & malicious noise → receive little to nothing

➡ Require: Evaluate each user's contributions, Identify client quality

Solution: Quality-aware data valuation in ML context

- Valuate the data quality
- Quality in ML model: data's contribution to the model performance

$$Acc\big(M(D,i)\big) - Acc\big(M(D)\big) = \ ?$$

1.3 Shapley Value (SV)

$$\phi(i) = c \sum_{S \subset N/i} \frac{\left[v\left(M_{(S \cup i)}\right) - v\left(M_S\right)\right]}{\binom{N-1}{|S|}}$$

$O(2^n)$

$$v(c) = \begin{cases} 80, & \text{if } c = \{A\} \\ 56, & \text{if } c = \{B\} \\ 70, & \text{if } c = \{C\} \\ 80, & \text{if } c = \{A, B\} \\ 85, & \text{if } c = \{A, C\} \\ 72, & \text{if } c = \{B, C\} \\ 90, & \text{if } c = \{A, B, C\} \end{cases}$$

π	δ_π^G
(A, B, C)	$(80, 0, 10)$
(A, C, B)	$(80, 5, 5)$
(B, A, C)	$(24, 56, 10)$
(B, C, A)	$(18, 56, 16)$
(C, A, B)	$(15, 5, 70)$
(C, B, A)	$(18, 2, 70)$
ϕ	$(39.2, 20.7, 30.2)$

Properties

- Group rationality

  $\phi(D) = \sum_{i \in D} \phi(i)$

- Symmetry

  $\forall S \subseteq N \backslash \{i, j\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $i = j$

- Null player

  $\forall S \subseteq N \backslash \{i\}$,v(S $\cup \{i\}) = v(S)$,then $i = 0$.

- Additivity

  $\forall i \in N, \phi(v_1 + v_2, i) = \phi(v_1, i) + \phi(v_2, i)$

1.4 Existing works

01

Directly using Shapley value in FL

- Fair! But:
- Missing FL order effect
- Incur communication costs

02

Round calculation

- Solve the order effect problem! But:
- Still get the value after the overall training
- Purely adding round SV $\neq$ overall SV

03

Incentivize once at the end

- Simple! But:
- Long waiting time also hurts the incentive
- Not sufficient to capture quality changes

04

Using model aggregation to replace repeated retraining

- Good idea!
- We will use it!

05

Not efficient enough to scale the FL system to massively distributed users.

1.5 Our contributions

- **A real-time incentive payoff scheme**
  - Maximize incentive effectiveness
  - Capture user quality changes timely
  - Prove fairness in both per-round and overall FL framework
- **Novel clustering-based approximation method**
  - Keep computing costs under control
- **Data valuation-inspired federated aggregation optimization**
  - Gain better global model in given number of rounds

Content

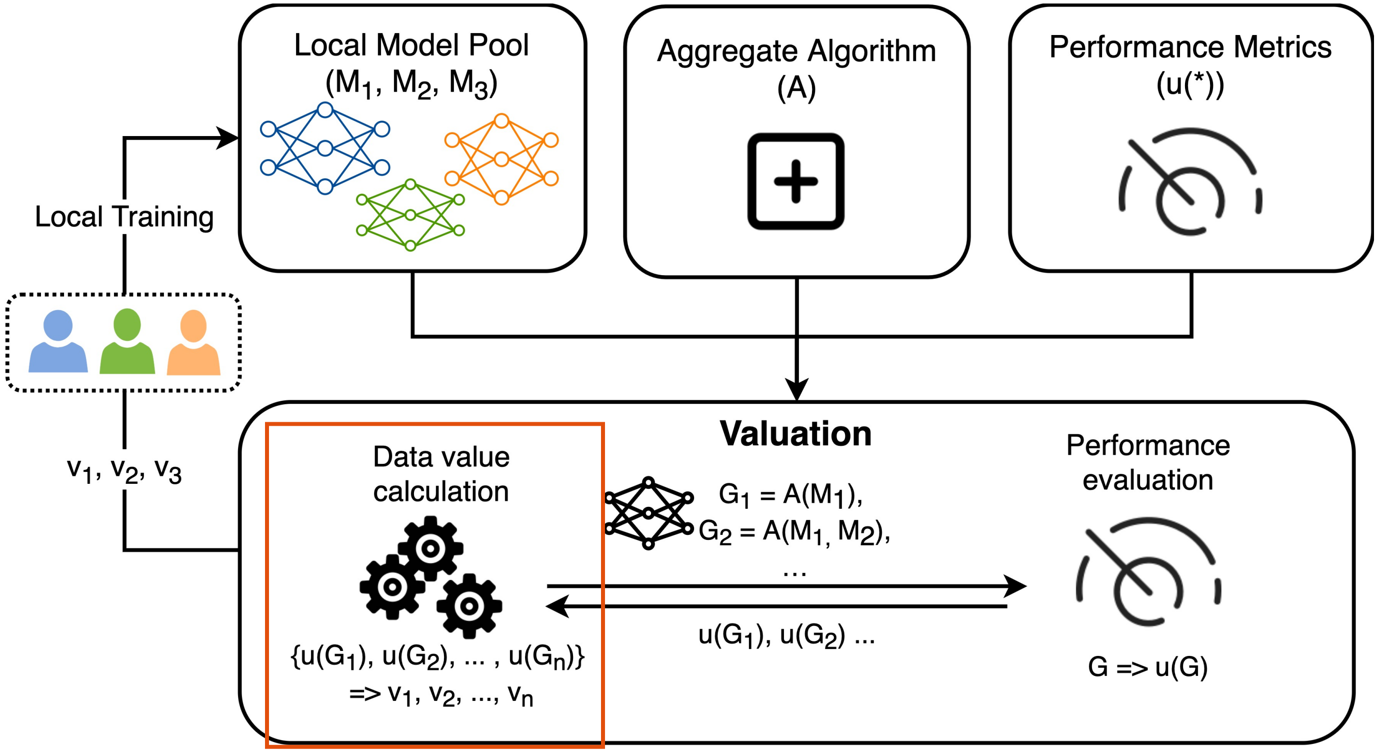
2.1 Round-based Data Valuation (RDV) with Shapley value

Model retraining ❌

Model aggregation ✓

In each round: calculate model gradient and use gradient descent to update global model

➡ Centered calculation
Resolve communication cost



Shapley value flow

2.1 Round-based Data Valuation (RDV) with Shapley value

Calculate SV in the fly of training

➡ Capture user quality changes

Distribute in real-time

➡ Reduce client waiting time

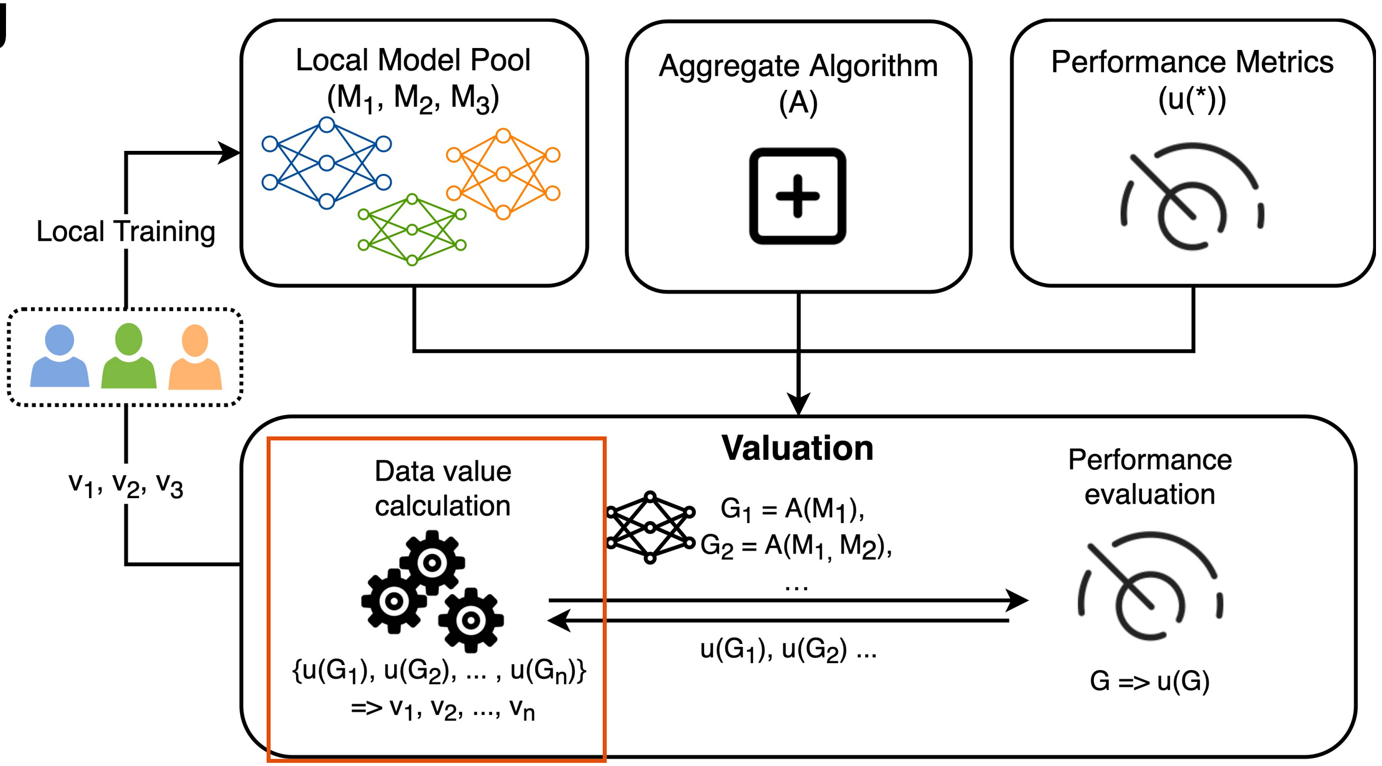
Decompose overall fairness to round fairness

➡ Group group rationality

Symmetry

Null player

Additivity



Shapley value flow

2.2 Sampling-based estimations

(A, B, C)
(A, C, B)
(B, A, C)
(B, C, A)
(C, A, B)
(C, B, A)

→

(A, B, C) → $u1$ for A, B, C
(A, C, B) → $u2$ for A, B, C
(B, A, C) → $u3$ for A, B, C

01 **K-subset stratified approximation (K-subset DV)**

02 **Truncated Monte-Carlo Sampling approximation (TMC-DV)**

$$\boldsymbol{O(nlogn)}$$

2.2 Clustering-based estimations

Clustering-based data valuation (CDV)

Group the local updates

$$||M_1 - M_2||_{\cos} = 1 - < m_1, m_2 > / \sqrt{|m_1|\,|m_2|}$$

⬇

Perform calculation in the unit of clusters

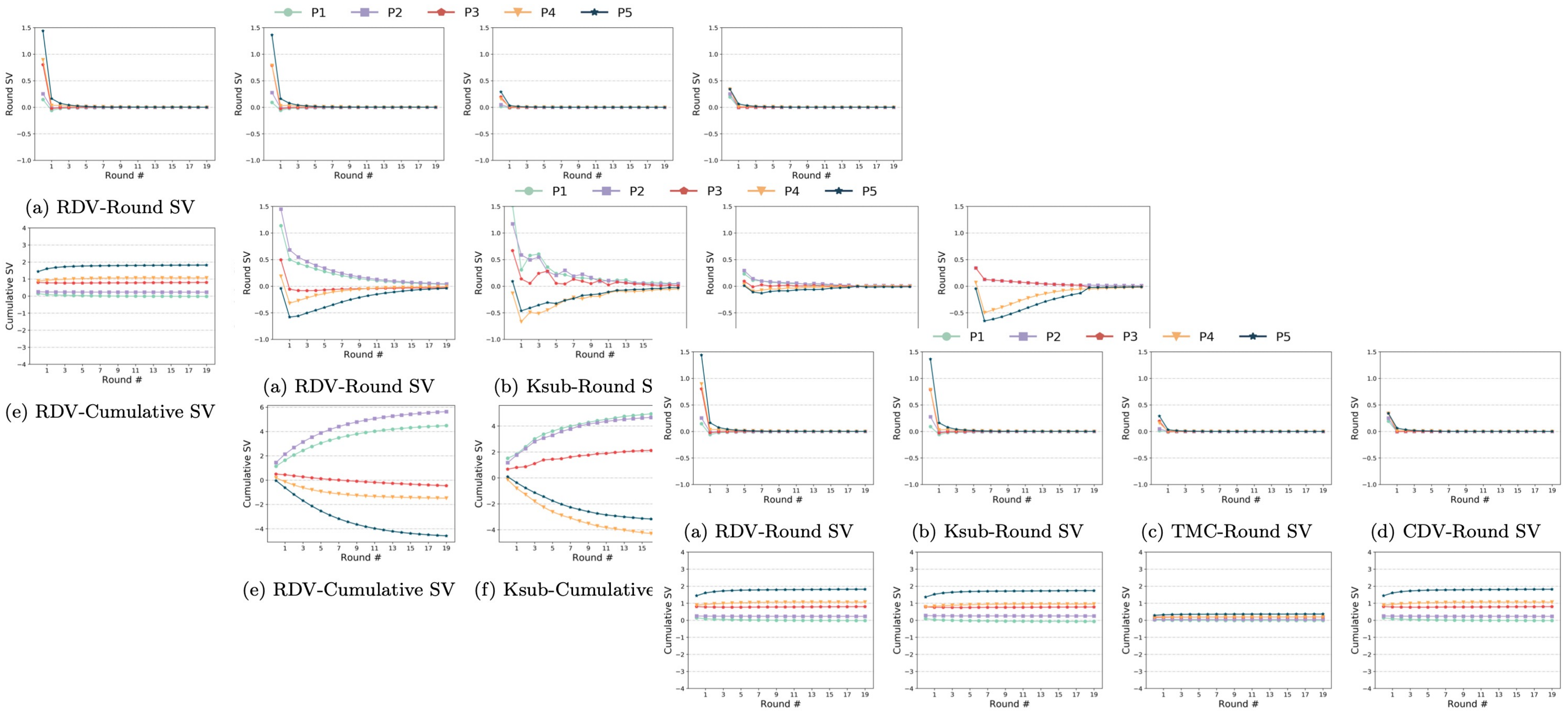
⬇

Assign the cluster value to each member equally

$$O(2^k)\ (k = cluster\ number)$$

Tradeoff: valuation accuracy vs. efficiency

3.1 Experiment result – effectiveness (1)



(a) RDV-Round SV

(e) RDV-Cumulative SV

(a) RDV-Round SV

(b) Ksub-Round S

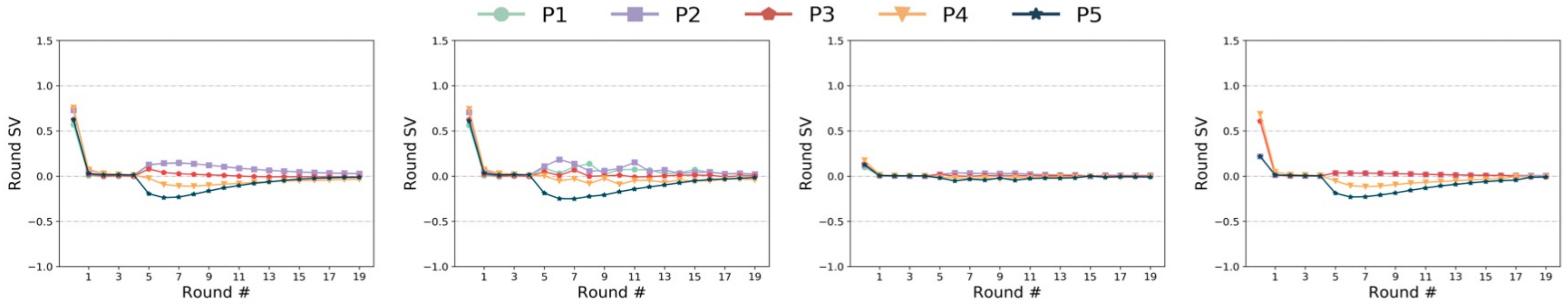
(e) RDV-Cumulative SV

(f) Ksub-Cumulative

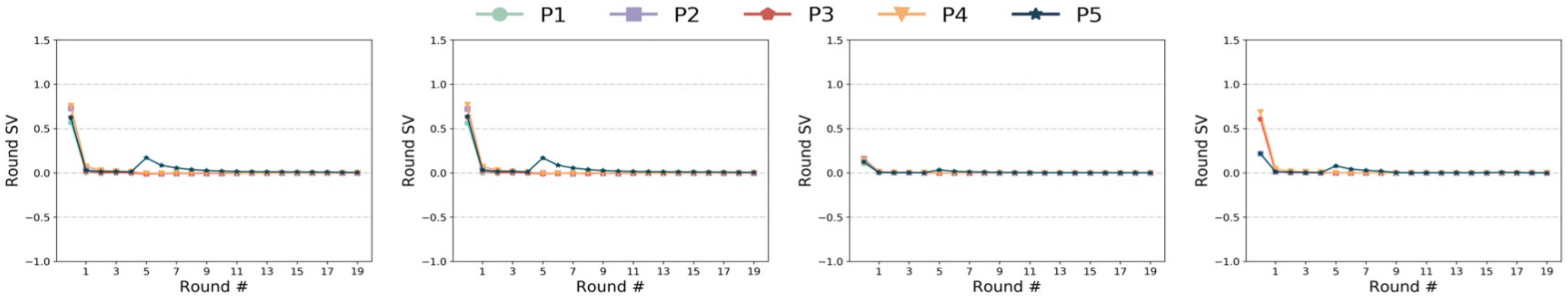
(a) RDV-Round SV (b) Ksub-Round SV (c) TMC-Round SV (d) CDV-Round SV

(e) RDV-Cumulative SV (f) Ksub-Cumulative SV (g) TMC-Cumulative SV (h) CDV-Cumulative SV

3.1 Experiment result – effectiveness (2)

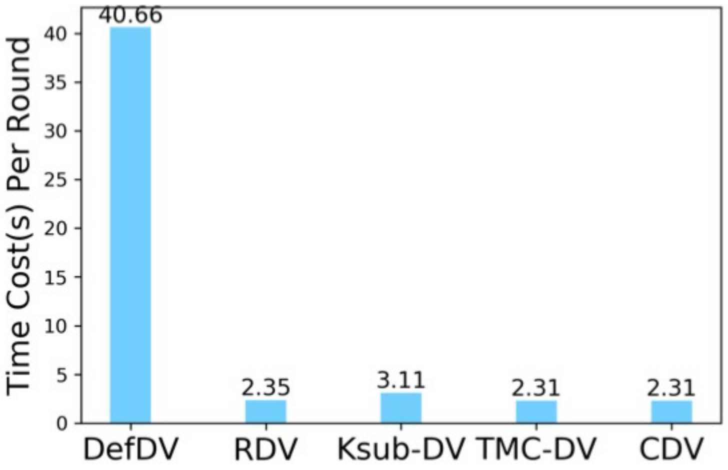


(a) RDV (b) Ksub (c) TMC (d) CDV

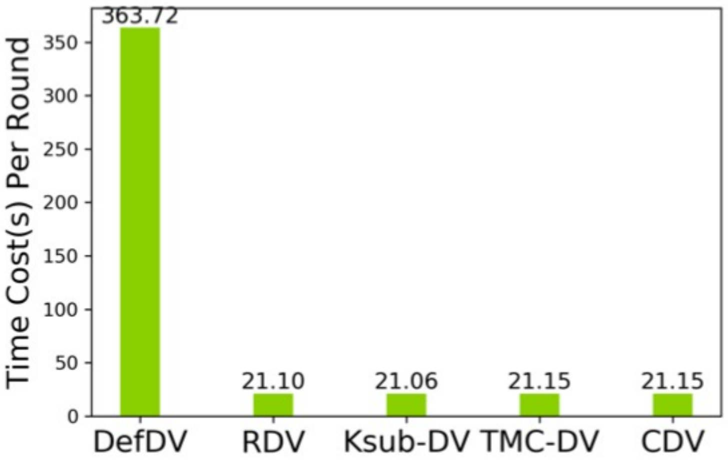


(a) RDV (b) Ksub (c) TMC (d) CDV

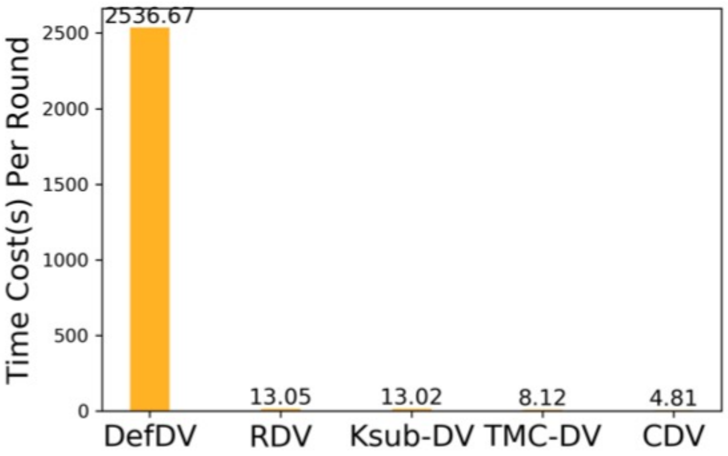
3.2 Experiment result – efficiency



(a) 5 p + 1000 local data



(b) 5 p + 10000 local data



(c) 10 p + 1000 local data

2.3 Data valuation-based selective aggregation

Basic Idea:
Average aggregation

Problem:

Noise insertion → Hurt global model

Use data valuation results →

1. Positive-Only Strategy

Only choose those who have positive values

2. Positive-Weighted Strategy

- Weights are based on values
- Weighted aggregation

2.4 RANSAC-selective aggregation

Main Idea:

- No exact Shapley value calculation
- Find the optimal participant set to update the global model by iteration

Process:

Repeat $\boldsymbol{k}$ times: Random sample n participants – Aggregate and Test – Record the sample value to each member

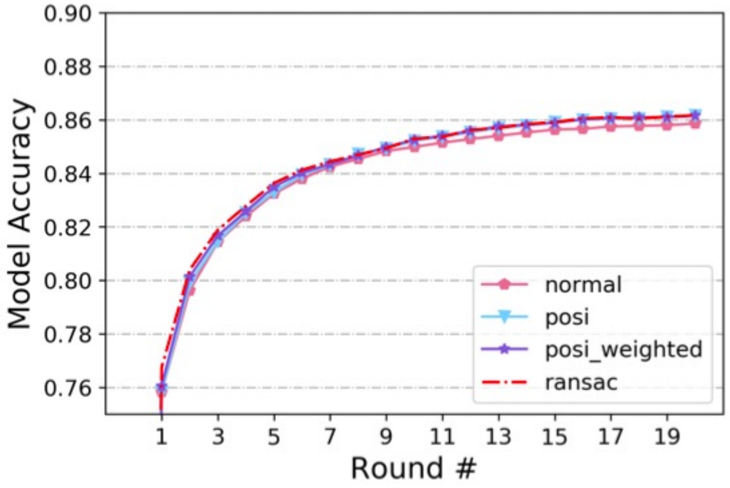
↓

Accumulate

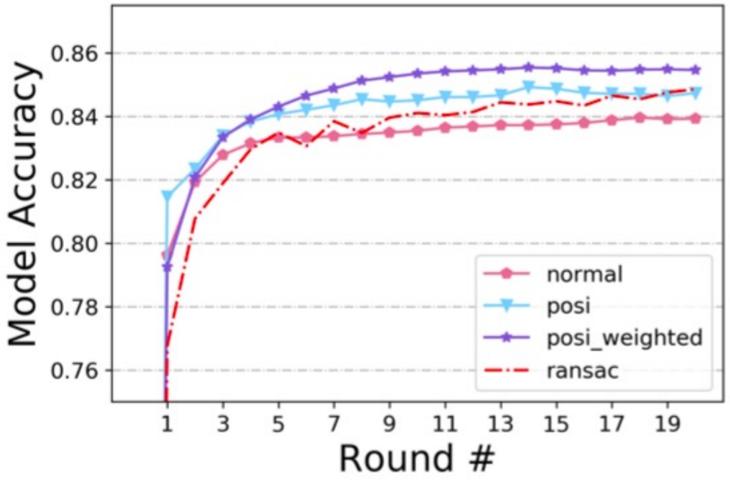
↓

Selection

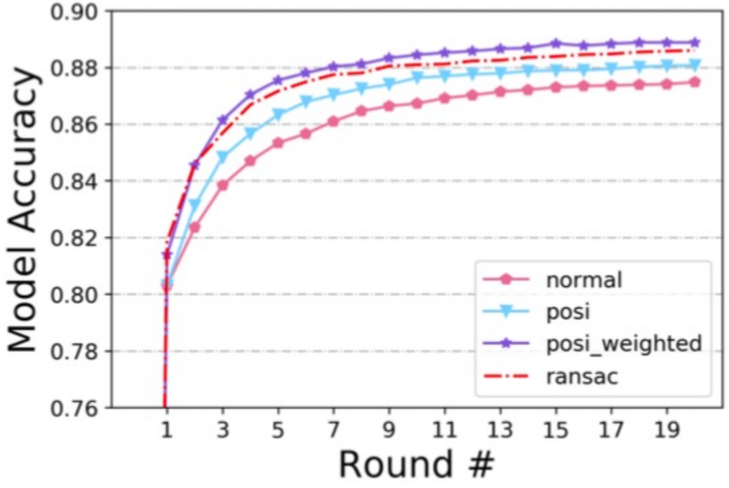
3.3 Experiment result – effectiveness & robustness



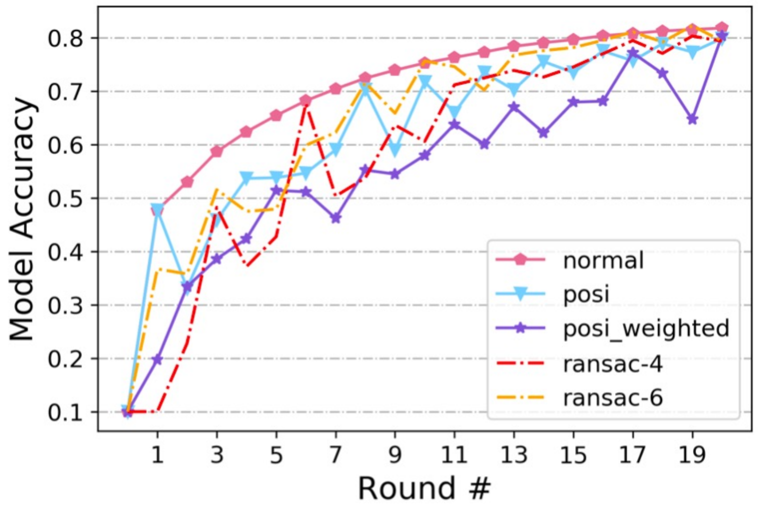
(a) OD environemnt



(b) ND environment



(c) UD environment



(d) Non-IID environment

Content

4 Conclusion

- Propose a FL-specific round data valuation approach (RDV) and their estimations to serve as FL incentive scheme.
- Suggest data valuation-inspired federated optimizations.
- A starting point in data valuation-based incentive scheme, will go on……

Thank You!

Q & A