针对受损源的联邦学习的性能加权

由于联邦学习对各种数据损坏攻击的敏感性，证明了局部权重的标准全局聚集方案在存在被破坏的客户端的情况下是低效的。为了缓解这个问题，我们提出了一类面向任务的基于性能的方法，这些方法在分布式验证数据集上计算，目标是检测和缓解被破坏的客户端。本文构造了一个基于几何平均的鲁棒权重聚合方案，并证明了它在随机标签洗牌和有针对性的标签翻转攻击下的有效性。

使用每个学习者的本地验证数据集作为测试平台来评估本地模型的性能，并测量它们在社区模型中的权重。实验证明，即使大多数训练和验证数据被破坏，我们的加权方案对数据中毒攻击也是鲁棒的。

DVW:

每个学习者将其本地数据集分成两个不相交的数据集，一个训练数据集和一个验证数据集（5%），并在整个联合执行中保留验证数据集，用于评估其他学习者的模型。由于联盟可能由具有不同数据量的学习者组成，并且每个类的训练样本的数量可能不均衡，因此通过分层采样来组装验证数据集。

在聚集局部模型之前，控制器对照每个参与学习者的验证数据集评估每个局部模型，以便确定每个单独模型的性能加权因子，控制器将每个学习器的本地模型发送到每个其他学习器的评估器服务，并累积来自所有服务的相应评估度量，控制器将来自每个评估器服务的混淆矩阵组合成一个累积混淆矩阵。

绩效加权分数：

微观平均精度：

$$DVW_{acc}^{\mu} = \frac{TP_C}{\#Examples}$$

宏观平均精度：

$$DVW_{acc}^{M} = \frac{\sum_i^C a_i}{C}$$

几何平均值：

$$DVW^{GMean} = \sqrt[C]{\Pi_i^C a_i}$$

Algorithm 1 Performance Weighting.

Controller executes:

for $t = 0, \ldots, T-1$ **do**

for each learner $k \in N$ **do**

$w_k = \text{CLIENTOPT}(w_c, \epsilon)$

$p_k = \text{EVAL}(w_k)$

$w_c = \sum_{k=1}^{N} \frac{p_k}{\mathcal{P}} w_k$ with $\mathcal{P} = \sum_{k}^{N} p_k$

CLIENTOPT(w_t, ϵ)**:**

$\mathcal{B} \leftarrow$ Split $\epsilon * D_k^T$ in batches of size β

for $b \in \mathcal{B}$ **do**

$w_{t+1} = w_t - \eta \nabla F_k(w_t; b)$

Return w_{t+1}

EVAL(w)**:**

$CM = 0_{C,C}$ {Confusion matrix CxC}

for each learner $k \in N$ **in parallel do**

$CM = CM + \text{EVALUATOR}_k(w)$

SCORE $= fn(CM)$

Return SCORE

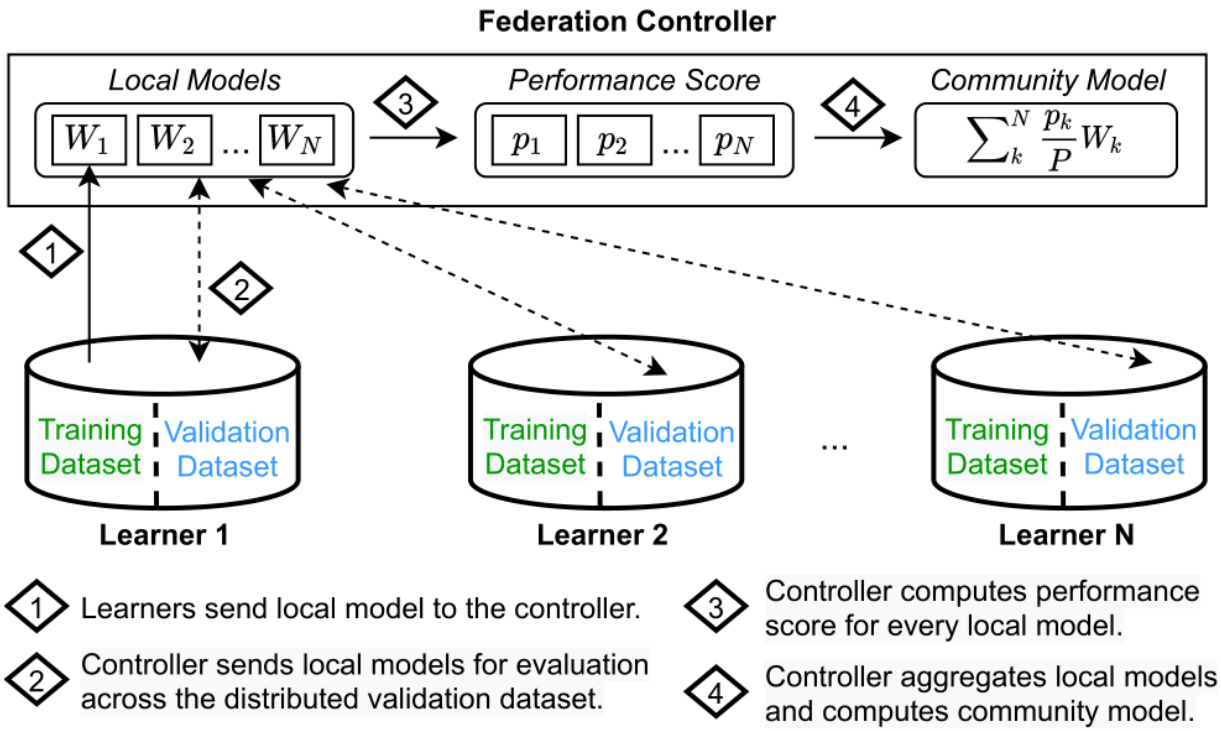
训练过程：

(1) 学习者在他们的本地数据集上进行本地训练。

(2) 控制器接收训练的模型并将它们发送给每个参与学习者的评估服务进行评估。

(3) 控制器计算每个本地模型的性能分数。

(4) 控制器聚集本地模型并计算全局模型。



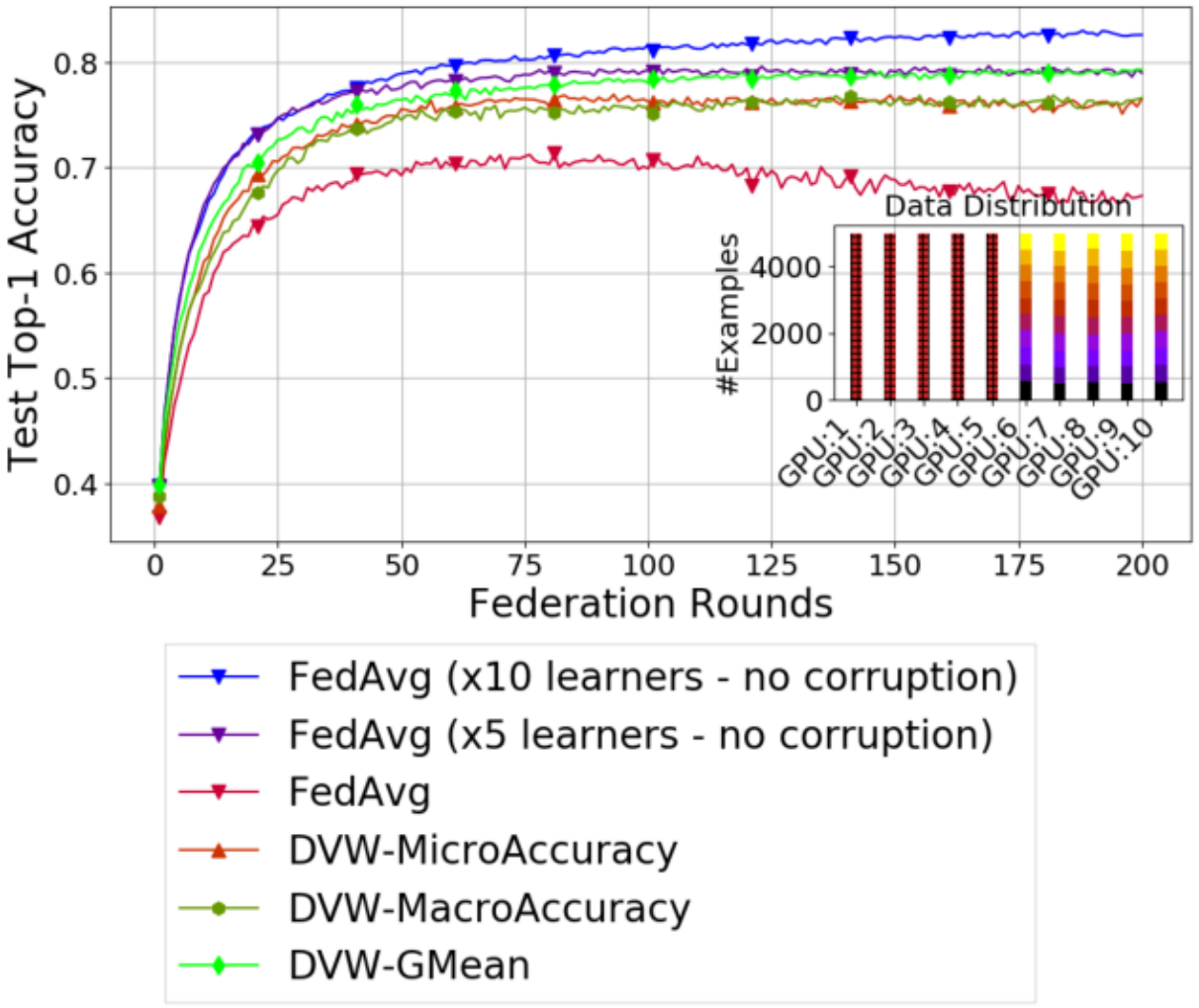
攻击：

主要探索了两种不同的数据中毒攻击，标签洗牌和目标标签翻转。

标签混洗：

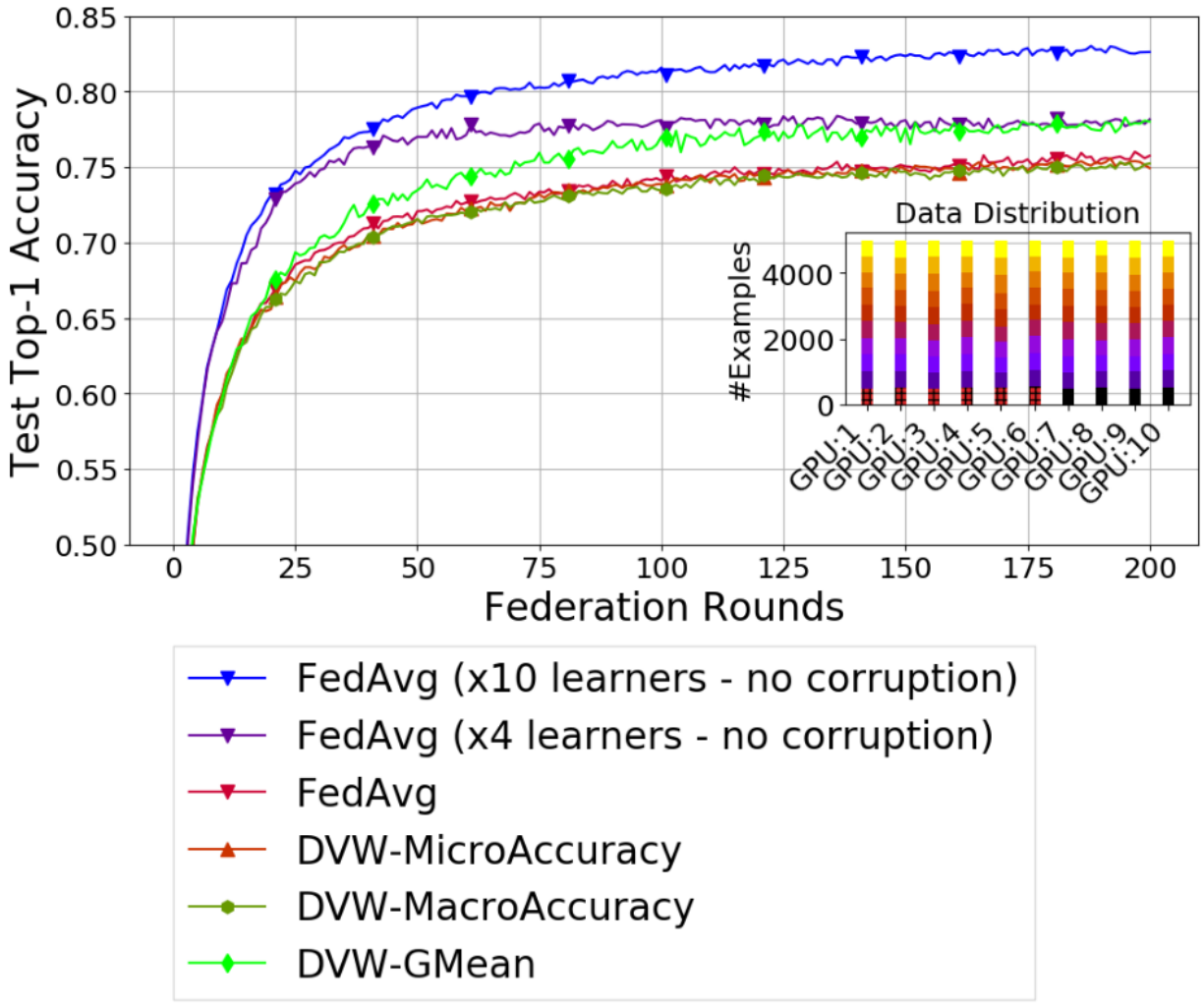
标签混洗的情况下，学习器的本地数据集的标签被随机打乱。

表明标签混洗的影响只是训练时间的适度增加。



目标标签翻转：

目标标签翻转攻击指的是被破坏的学习器将对应于特定源类的示例的标签翻转到目标类的情况。



结论：

实验证明基于几何平均的聚集方案对于随机标签洗牌和目标标签翻转具有更好的弹性。本方案仍然受益于部分损坏的站点，在某些情况下，比仅仅将它们排除在联盟之外获得更好的结果。