Data Appraisal Without Data Sharing

Mimee Xu[*]
New York University
mimee@nyu.edu

Laurens van der Maaten & Awni Hannun
Facebook AI Research, New York
{lvdmaaten,awni}@fb.com

Abstract

One of the most effective approaches to improving the performance of a machine-learning model is to acquire additional training data. To do so, a model owner may seek to acquire relevant training data from a data owner. Before procuring the data, the model owner needs to appraise the data. However, the data owner generally does not want to share the data until after an agreement is reached. The resulting Catch-22 prevents efficient data markets from forming. To address this problem, we develop data appraisal methods that do not require data sharing by using secure multi-party computation. Specifically, we study methods that: (1) compute parameter gradient norms, (2) perform model fine-tuning, and (3) compute influence functions. Our experiments show that influence functions provide an appealing trade-off between high-quality appraisal and required computation.

1 Introduction

In many real-world applications, the quality of machine-learning models depends heavily on the amount of data that is used to train the models. As a result, model developers may want to acquire (additional) training data from external data owners. For example, a weather forecaster may improve their models by procuring additional weather satellite images from a satellite company. This requires the model owner to appraise data: they need to estimate the utility they will get from the data, so they can determine which data is worth procuring. Such data appraisal is non-trivial because the value of data to a model owner depends on many factors, including what data the model owner already has, the complexity of their model, the data distribution on which they seek to perform predictions, *etc.*

Ideally, the model owner would: (1) re-train their model with and without the data to be appraised and (2) measure the accuracy gain on the test set that resulted from using the additional training data. However, performing data appraisal requires the data owner to share their data with the model owner before the model owner has acquired the data. Data owners may be unwilling to do this, which leads to a Catch-22 that can prevent an efficient data market from forming.

To address this problem, we develop techniques that perform data appraisal without requiring data sharing. Specifically, we use secure multi-party computation (MPC) to develop and evaluate three methods for data appraisal without data sharing: (1) a method that computes parameter gradient norms, (2) a method that performs model fine-tuning, and (3) a method that computes influence functions. The results of our experiments show that computing influence functions via secure MPC allows for high-quality data appraisal while requiring relatively limited amounts of additional computation.

2 Problem Setting

We assume a *model owner*, who is developing a machine-learning model with parameters θ, and a *data owner*, who possesses the dataset to be appraised, $\mathcal{D}_\text{a}$. The model owner has a seed training

[*]Research performed as an intern at Facebook AI Research, New York

NeurIPS Workshop on Privacy-Preserving Machine Learning (PPML 2020).

dataset, $\mathcal{D}_{\text{tr}}$, and a test set, $\mathcal{D}_{\text{te}}$, to evaluate their model. The model owner wishes to determine the utility gain they would obtain by acquiring $\mathcal{D}_{\text{a}}$ and training their model on $\mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{a}}$.

The model owner computes the model parameters $\hat{\theta}$ by minimizing the regularized empirical risk on the seed training dataset, $\mathcal{D}_{\text{tr}}$:

$$\hat{\theta} = \arg\min_{\theta} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{tr}}} L(\boldsymbol{x}, y; \theta) + \lambda\|\theta\|_2^2. \quad (1)$$

We assume that the loss $L(\cdot)$ is twice differentiable and convex in θ. Given an additional dataset, $\mathcal{D}_{\text{a}}$, we can compute the new optimal parameters, θ^*, by minimizing the regularized empirical risk on dataset $\mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{a}}$ instead of dataset $\mathcal{D}_{\text{tr}}$. We define the utility of the dataset $\mathcal{D}_{\text{a}}$ for the model owner as:

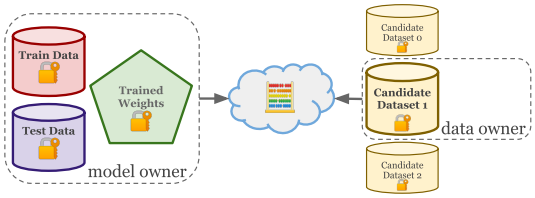

Figure 1: In secure multi-party computation, the model owner and data owner encrypt their data prior to the appraisal. The appraisal is performed privately, and its result is opened up to both parties.

$$U(\mathcal{D}_{\text{a}}) = \frac{1}{|\mathcal{D}_{\text{te}}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{te}}} L(\boldsymbol{x}, y; \theta^*) - L(\boldsymbol{x}, y; \hat{\theta}). \quad (2)$$

The goal of data appraisal is to approximate this utility without requiring the model and data owners to share the model parameters, $\hat{\theta}$, or any of the datasets $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{te}}$, and $\mathcal{D}_{\text{a}}$. We develop three methods that implement an appraisal function $f(\mathcal{D}_{\text{a}})$ designed to approximate $U(\mathcal{D}_{\text{a}})$. A good appraisal function $f(\cdot)$ produces the same ranking over datasets as the utility function $U(\cdot)$.

Threat Model. We assume a passively secure threat model. Both the model and data owners are *honest-but-curious*. The parties follow the MPC protocol but should not be able to learn anything from the data observed. We assume the appraisal of the dataset is revealed to both parties, and that the parties accept the associated information leakage. If such information leakage is unacceptable, the appraisal value can be kept secret and a single bit representing the acquisition decision can be revealed. This may require the model owner to pre-define a threshold on the value of $f(\mathcal{D}_{\text{a}})$.

We also assume that metadata about the dataset to be appraised is available to both parties. This metadata includes the number of data samples, their dimensionality, and the number of classes. Relevant metadata may also include details on data type, data encoding, label encoding, *etc.*

3 Data Appraisal Without Data Sharing

We implement and study three private data appraisal methods, using the secure MPC implementations in CrypTen [7]. Secure MPC is well-suited for our purpose, because it allows parties to jointly evaluate functions on their combined data without revealing that data (which includes model parameters) or any intermediate values in the function evaluation. The appraisal methods we study are based on: (1) the norm of the parameter gradient induced by the data, (2) finetuning of the model using Stochastic Gradient Descent (SGD), and (3) evaluating influence functions.

Norm of Parameter Gradients. Data can be appraised via the norm of the loss gradient that dataset $\mathcal{D}_{\text{a}}$ induces in the model parameters $\hat{\theta}$:

$$f_{\text{gn}}(\mathcal{D}_{\text{a}}) = \left\| \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{a}}} \nabla_\theta L\left(\boldsymbol{x}, y; \hat{\theta}\right) \right\|_2. \quad (3)$$

An advantage of this approach is that it only requires the parameters $\hat{\theta}$ and the additional data $\mathcal{D}_{\text{a}}$, but not a test set. However, the gradient norm may not always approximate utility well as a result: for example, it can be large when $\mathcal{D}_{\text{a}}$ contains unfamiliar but useless or even harmful data.

Model Finetuning. Finetuning a model on $\mathcal{D}_{\text{a}} \cup \mathcal{D}_{\text{tr}}$ can approximate data utility arbitrarily well:

$$f_{\text{ft}}(\mathcal{D}_{\text{a}}) = \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{te}}} L(\boldsymbol{x}, y; \hat{\theta}_{\text{ft}}) - L(\boldsymbol{x}, y; \hat{\theta}), \quad (4)$$

Table 1: Correlation ρ of appraised values and data utility under varying amounts of label noise. Finetuning runtimes are limited to $1\times$, $4\times$ and $16\times$ of influence runtime, each benchmarked on the best performances under three learning rates: 0.001, 0.1, and 10. Higher is better.

	Gradient norm	**Finetuning** $(1\times)$	**Finetuning** $(4\times)$	**Finetuning** $(16\times)$	**Influence**
$\boldsymbol{\rho}$	-1.00	0.61, 0.95, 0.96	0.58, 1.0, 0.59	0.72, 1.0, 0.88	0.96

where $\hat{\theta}_{\text{ft}}$ are the parameters after a fixed number of SGD updates on $\mathcal{D}_{\text{a}} \cup \mathcal{D}_{\text{tr}}$ seeded with $\hat{\theta}$. A downside of model finetuning is that it is computationally very intensive when implemented via MPC. Moreover, successful SGD optimization often requires careful tuning of hyper-parameters, which is difficult in secure MPC settings in which inspection of the training loss is not possible.

Influence Functions. The influence function $\mathcal{I}(\boldsymbol{x}, y)$ associates a training sample with the change in the model parameters under an infinitisemal up-weighting of that sample in the risk [2, 8]. Defining the Hessian $\boldsymbol{H}_{\hat{\theta}} = \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{tr}}} \nabla^2_{\theta} L(\boldsymbol{x}, y, \hat{\theta})$, the influence of sample $(\boldsymbol{x}, y)$ is given by:

$$\mathcal{I}(\boldsymbol{x}, y) = -\boldsymbol{H}_{\hat{\theta}}^{-1} \nabla_{\theta} L(\boldsymbol{x}, y, \hat{\theta}). \tag{5}$$

This function allows us to approximate the change in $\hat{\theta}$ for each sample in $\mathcal{D}_{\text{a}}$. In turn, we can use this approximation to assess the influence of $(\boldsymbol{x}', y')$ on the test loss of $(\boldsymbol{x}, y)$ via the chain rule:

$$L(\boldsymbol{x}, y; \theta^*) - L(\boldsymbol{x}, y; \hat{\theta}) \approx -\nabla_{\theta} L(\boldsymbol{x}, y; \hat{\theta})^{\top} \mathcal{I}(\boldsymbol{x}', y'). \tag{6}$$

Using these observations, we define the influence-based appraisal function to be:

$$f_{\text{if}}(\mathcal{D}_{\text{a}}) = -\frac{1}{|\mathcal{D}_{\text{a}}| \cdot |\mathcal{D}_{\text{te}}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{te}}} \sum_{(\boldsymbol{x}',y')\in\mathcal{D}_{\text{a}}} \nabla_{\theta} L(\boldsymbol{x}, y; \hat{\theta})^{\top} \boldsymbol{H}_{\hat{\theta}}^{-1} \nabla_{\theta} L(\boldsymbol{x}', y'; \hat{\theta}). \tag{7}$$

We note that the inverse Hessian is fixed; it only needs to be computed once and can be computed in the clear by the model owner. Hence, only the loss gradient and an inner product in $\mathbb{R}^d$ need to be computed using secure MPC (for each sample in $\mathcal{D}_{\text{a}}$). This makes evaluating $f_{\text{if}}(\mathcal{D}_{\text{a}})$ efficient.

4 Experiments

We evaluate the efficacy of our data appraisal methods in two scenarios: (1) a scenario in which the utility of the data varies because of label noise in that data and (2) a scenario in which the utility varies because the data distribution does not match the distribution that the model owner is interested in. In all experiments, the models are L2-regularized logistic regressors. We train and evaluate the model on classification problems using the MNIST and CIFAR-10 datasets: on MNIST, we classify ten digits, and on CIFAR-10, we distinguish planes from cars. Prior to evaluating the appraisal functions on $\mathcal{D}_{\text{a}}$, we train the model with L-BFGS on the seed training set $\mathcal{D}_{\text{tr}}$ until convergence.

Label Noise. In our first set of experiments, we vary the utility of the dataset $\mathcal{D}_{\text{a}}$ by introducing label noise. In particular, we use 1% of the MNIST training data as $\mathcal{D}_{\text{tr}}$. The remaining training data is split into 10 candidate datasets $\mathcal{D}_{\text{a}}^{(p)}$ with $p = 1, \ldots, 10$. For each of the candidate sets $\mathcal{D}_{\text{a}}^{(p)}$, we randomly flip labels 1 and 7 with probability $p/10$. We evaluate models on $\mathcal{D}_{\text{te}}$.

Table 1 presents the correlation ρ of the label-noise probabilities with the appraisal value for all three methods, including under three finetuning learning rates: 0.001, 0.1, and 10. The correlations are high for the model finetuning and influence function methods, suggesting that the appraisal value accurately captures data utility. Surprisingly, the gradient norm method predicts lower utility for datasets with less label noise. This effect is also illustrated by Figure 2, which presents appraisal value (y-axis; note that the units vary per method) as a function of datasets under our two sets of experiments: label noise (x-axis) on MNIST (top) and data imbalance on CIFAR-10 (bottom; see next paragraph for details). Figure 3 presents the total runtime of each appraisal function, separating the encrypted from the plaintext runtimes under MNIST setup, assuming 16 steps of full-batch gradient descent for fine-tuning. Figure 4 presents the effect of finetuning hyperparameters on the correlation of appraisal with utility (top) and runtime (bottom). Overall, the results suggest that influence functions provide an appealing trade-off between high predictive value of data utility (even for large datasets) and runtime of the computation, while not requiring any hyperparameter optimization.

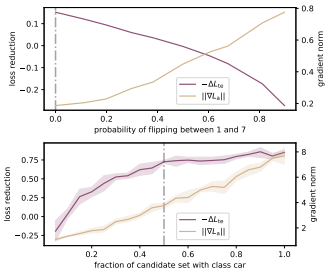

Figure 2: Gradient norm appraisal and test loss reduction as a function of MNIST label noise (top) and CIFAR-10 class balance (bottom).

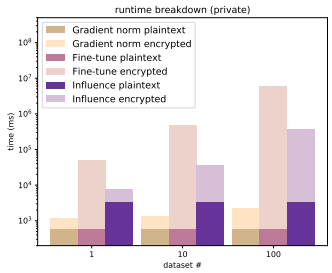

Figure 3: Runtime (log scale) of the plaintext and encrypted part of each computation for all three data appraisal methods.

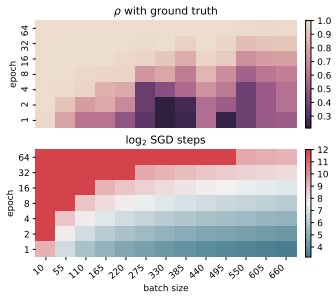

Figure 4: Correlation of appraisal with utility (top; purple is lower) and runtime (bottom; blue is faster) for finetuning hyperparameters.

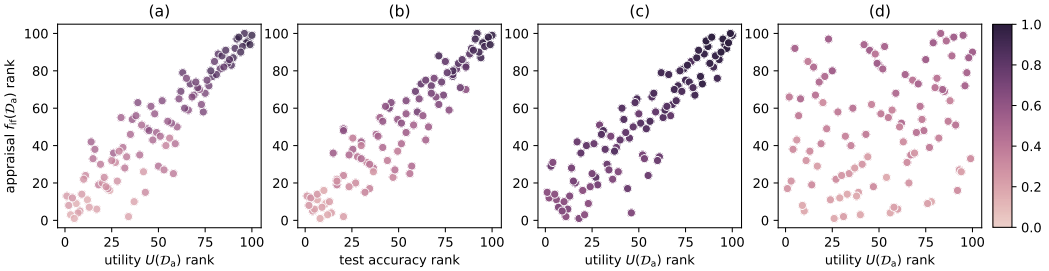

Figure 5: **Left:** Rank of influence-based appraisal value $f_{\text{if}}(\mathcal{D}_{\text{a}})$ (y-axis) as a function of the rank of the utility (a; $\rho = 0.923$) and the test accuracy (b; $\rho = -0.927$) on CIFAR-10. **Right:** Rank of $f_{\text{if}}(\mathcal{D}_{\text{a}})$ as a function of the rank of the utility on CIFAR-10 dataset for which the rate of cars is in the range $[0, 0.45]$ (c; $\rho = 0.908$) and in $[0.55, 1.0]$ (d; $\rho = 0.247$). Colors indicate the ratio of the undersampled class in $\mathcal{D}_{\text{a}}$.

Distribution Mismatch. In our second set of experiments, we focus on influence-based appraisal and study its efficacy under distribution mismatch. We simulate the mismatch between: (1) $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$ and (2) the candidate datasets $\mathcal{D}_{\text{a}}^{(p)}$ by varying the prior over classes. To do so, we construct a training set from CIFAR-10 with a 10:1 ratio of plane-to-car and a balanced test set with a 1:1 ratio of plane-to-car. We then construct 20 candidate datasets $\mathcal{D}_{\text{a}}^{(p)}$ of which exactly $(5 \cdot p)\%$ are planes and the remainder are cars, with $p = 1, \ldots, 20$. The candidate datasets are of size $|\mathcal{D}_{\text{a}}^{(p)}| = 440$. We repeat this process five times, sampling the datasets randomly each time.

Figure 5 shows scatter plots of: (a) the rank of the influence-based appraisal value, $f_{\text{if}}(\mathcal{D}_{\text{a}})$, of each of the 5×20 candidate datasets and (b) the rank of the utility or test accuracy of those datasets (see caption for details). The experimental results show that the influence-based appraisal value correlates well with gains in utility. Indeed, the appraisal value allows the model owner to select a candidate dataset that closely resembles their desired distributions in most situations. However, when zooming in on different ranges of class ratios, the influence-based appraisal value, $f_{\text{if}}(\mathcal{D}_{\text{a}})$, becomes less informative when the class ratio deviates far from that of both the training and testing datasets (c-d).

5 Conclusion

We have demonstrated that it is possible for a model owner to appraise another party's data without requiring any data (or model) sharing between the two parties. We studied three data-appraisal methods that operate in this setting by leveraging secure MPC techniques. Our empirical results suggest that appraising data using influence function leads to accurate valuations in many scenarios, while requiring limited computation and no hyperparameter optimization.

We aim to extend this work by studying data appraisal without data sharing on real-world datasets using more complex models. We also plan to improve the data appraisal methods we considered

by including a differentially private mechanism to bound the information that is leaked when the appraisal value itself is released.

Acknowledgments

The authors thank Brian Knott and Alex Melville for helpful discussions.

References

[1] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[2] R. D. Cook and S. Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

[3] A. Ghorbani and J. Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, 2019.

[4] R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147, 2019.

[5] C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning*, 2020.

[6] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176, 2019.

[7] B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten. Crypten: Secure multi-party computation meets machine learning. 2020.

[8] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017.

[9] C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. *ACM Transactions on Database Systems (TODS)*, 39(4):1–28, 2014.

[10] L. S. Shapley. A value for n-person games. Technical report, Rand Corp Santa Monica CA, 1952.

[11] T. Song, Y. Tong, and S. Wei. Profit allocation for federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2577–2586. IEEE, 2019.

[12] G. Wang, C. X. Dang, and Z. Zhou. Measure contribution of participants in federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2597–2604. IEEE, 2019.

[13] T. Wang, J. Rausch, C. Zhang, R. Jia, and D. Song. A principled approach to data valuation for federated learning, 2020.

A Related Work

Assessing the impact of data to a statistical model is a long studied subject. A natural method is to measure the effect of the data under leave-one-out training–also known as Cook's distance in linear regression [1]. Influence functions can be used to efficiently approximate leave-one-out training [2, 8]. Much prior work exists that uses influence functions for various applications including interpretability [8], efficient cross-validation [4], and efficient training data removal [5]. However, little work exists that uses influence functions for data appraisal.

More recently, Shapley values [10] have been proposed as a more equitable method for data appraisal [3, 6]. A primary motivation of Shapley values over influence-based approaches is the invariance to the order of data aquisition. Instead, we focus on the case where the order of acquisition is important. Therefore, we opt for using leave-one-out and influence-based valuation techniques instead.

Prior work also considers private data appraisal with differential privacy [9] and federated learning [11, 12, 13]. The problem setting considered in these studies is different from the setting studied in our

work. The exchange of data in these applications is treated as foregone conclusion, with the private pricing serving as a mechanism to incentivize more data contributions. In contrast, our protocol allows the appraisal to be performed in private prior to the transaction. This lets the model and data owners decide if it is worth engaging in the transaction based on the appraised value.