What is your data worth? Equitable Valuation of Data

Amirata Ghorbani and James Y. Zou - Department Electrical Engineering, Stanford University, CA, USA

28 Feb 2020

Presenter: Sanchit Sinha

https://qdata.github.io/deep2Read/

Motivation

- **Data is important:** in current state of the art classification models involving machine learning
- **Not all data created equal:** some data points are more useful in training the classifier - they have higher discriminative value
- **How to quantify the quality of data:** the paper explores this fundamental question regarding how to ascribe scores to data points wrt the usefulness in model training
- **Recognizing data as property:** laws are being created to start identifying data as a property of a person, and hence, there is intrinsic value in the data.
- **Compensating people who provide data:** people (especially in medical fields) volunteer to provide data. However, to motivate people to give data, there has to be some compensation which should depend on the quality of the data

Background

- Supervised Learning **ingredients**: training data (D), learning algorithm(A), and performance metric (V)
- **Assigning value** to data in the setting of supervised learning is not equal to assigning a universal to it - all learning algorithms depend on data
- **Shapley value**: In a cooperative game, there are n players $D = \{1,\ldots,n\}$ and a score function $v : 2^n \rightarrow R$ assigns a reward to each of 2^n subsets of players: v(S) is the reward if the players in subset $S \subseteq D$ cooperate.
- Shapely uniquely **divides the reward** for cooperation of all players V(D) such that each player would get an equitable share of the reward.

Related Work

- Shapley, L. S. A value for n-person games. Contributions to Theory Games 2, 307–317 (1953)
- Bounding the estimation error of sampling-based shapley value approximation. arXiv preprint arXiv:1306.4265 (2013)
- Steinhardt, J., Koh, P. W. W. & Liang, P. S. Certified defenses for data poisoning attacks. In Advances in Neural Information Processing Systems

Claim / Target Task

- Come up with a technique to assign scores to data points depending on how good their quality is (how well they help in training the model)
- Build a better valuation method than “leave one out” - the most intuitive way of assigning a score to the data point
- Viewing the supervised machine learning problem as a cooperative game: each source in the train data is a player, and the players work together through the learning algorithm A to achieve prediction score
- Speed up the number of computations required in the Data Shapely equation
- Applying to Domain Adaptation by concentrating on the good quality data

Why not LOO?

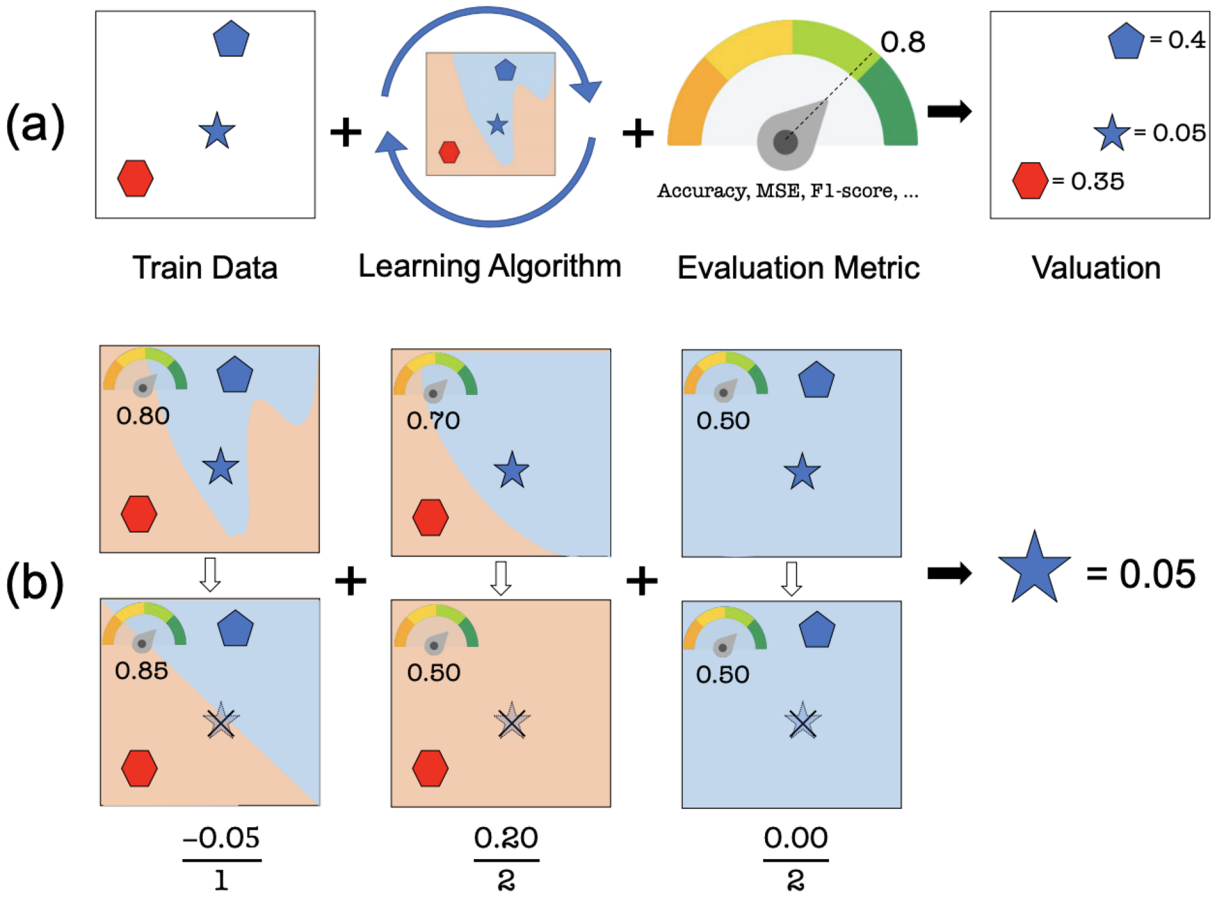
- Suppose a scenario where the training set contains n copies of each sample and consider using a KNN (with k<=n-1) to predict something using this data.
- If we remove any sample from the training set, the KNN's performance would not change since there are still n-1 identical copies of the removed sample in the dataset.
- In this case, the LOO would attribute 0 value to all data points.

Data Summary

Datasets used:

- ImageNet
- UK Biobank
- MNIST
- UPS
- Spam vs Email
- Skin Lesions

An Intuitive Figure Showing WHY Claim



Proposed Solution - Properties

- If (xi,yi) does not change the performance if it's added to any subset of the train data sources, then it should be given zero value. For all $S \subseteq D - \{i\}$, $V(S) = V(S \cup \{i\})$, then $\varphi i = 0$.
- If for data i and j and any subset $S \subseteq D - \{i, j\}$, we have $V(S \cup \{i\}) = V(S \cup \{j\})$, then $\varphi i = \varphi j$.
- When the overall performance score is the sum of separate performance scores, the overall value of a datum should be the sum of its value for each score: $\varphi i(V + W) = \varphi i(V) + \varphi i(W)$

$$\phi_i = C \sum_{S \subseteq D - \{i\}} \frac{V(S \cup \{i\}) - V(S)}{\binom{n-1}{|S|}}$$

Proposed Solution - TMC

Algorithm 1 Truncated Monte Carlo Shapley

Input: Train data $D = \{1, \ldots, n\}$, learning algorithm $\mathscr{A}$, performance score V
Output: Shapley value of training points: $\phi_1, \ldots, \phi_n$

Initialize $\phi_i = 0$ for $i = 1, \ldots, n$ and $t = 0$
while Convergence criteria not met **do**
  $t \leftarrow t + 1$
  π^t: Random permutation of train data points
  $v_0^t \leftarrow V(\emptyset, \mathscr{A})$
  for $j \in \{1, \ldots, n\}$ **do**
    if $|V(D) - v_{j-1}^t| <$ Performance Tolerance **then**
      $v_j^t = v_{j-1}^t$
    else
      $v_j^t \leftarrow V(\{\pi^t[1], \ldots, \pi^t[j]\}, \mathscr{A})$
    end if
    $\phi_{\pi^t[j]} \leftarrow \frac{t-1}{t}\phi_{\pi^{t-1}[j]} + \frac{1}{t}(v_j^t - v_{j-1}^t)$
  end for
end while

Proposed Solution - G Shapley

Algorithm 2 Gradient Shapley

Input: Parametrized and differentiable loss function $\mathscr{L}(.;\theta)$, train data $D = \{1,\ldots,n\}$, performance score function $V(\theta)$

Output: Shapley value of training points: $\phi_1,\ldots,\phi_n$

Initialize $\phi_i = 0$ for $i = 1,\ldots,n$ and $t = 0$

while Convergence criteria not met **do**

  $t \leftarrow t+1$

  π^t: Random permutation of train data points

  $\theta_0^t \leftarrow$ Random parameters

  $v_0^t \leftarrow V(\theta_0^t)$

  for $j \in \{1,\ldots,n\}$ **do**

    $\theta_j^t \leftarrow \theta_{j-1}^t - \alpha\nabla_\theta\mathscr{L}(\pi^t[j];\theta_{j-1})$

    $v_j^t \leftarrow V(\theta_j^t)$

    $\phi_{\pi^t[j]} \leftarrow \frac{t-1}{t}\phi_{\pi^{t-1}[j]} + \frac{1}{t}(v_j^t - v_{j-1}^t)$

  end for

end while

Experiments - Experiment 1a and 1b

- **Data Poisoning - Changing labels**:
  - Changing labels of a few data points in a train set.
  - Then assigning data Shapley values to all the data points - considering the ones which have the lowest values.
  - The experiments proved that the mislabelled points have the lowest Data Shapley values.
  - This opens the potential for correcting this data and possibly improving quality of low quality data
- **Data Poisoning - Corrupting data:**
  - Corrupting an image using noise.
  - Training the last layer of Inception v3 model
  - Again noisy images have least Shapley values

Experiments - Experiment 2 and 3

- **Adding data points of high Shapley values:**
  - Training a Random forest classifier on the predicted Shapley values
  - Only adding high Shapley value based on the RF classifier in the training set from a new set
- **Removing Low value points**:
  - Removing the points with lowest Shapley values

- **Domain Adaptation**: Removing low value points and training on a new task

Experimental Analysis - 1a and 1b

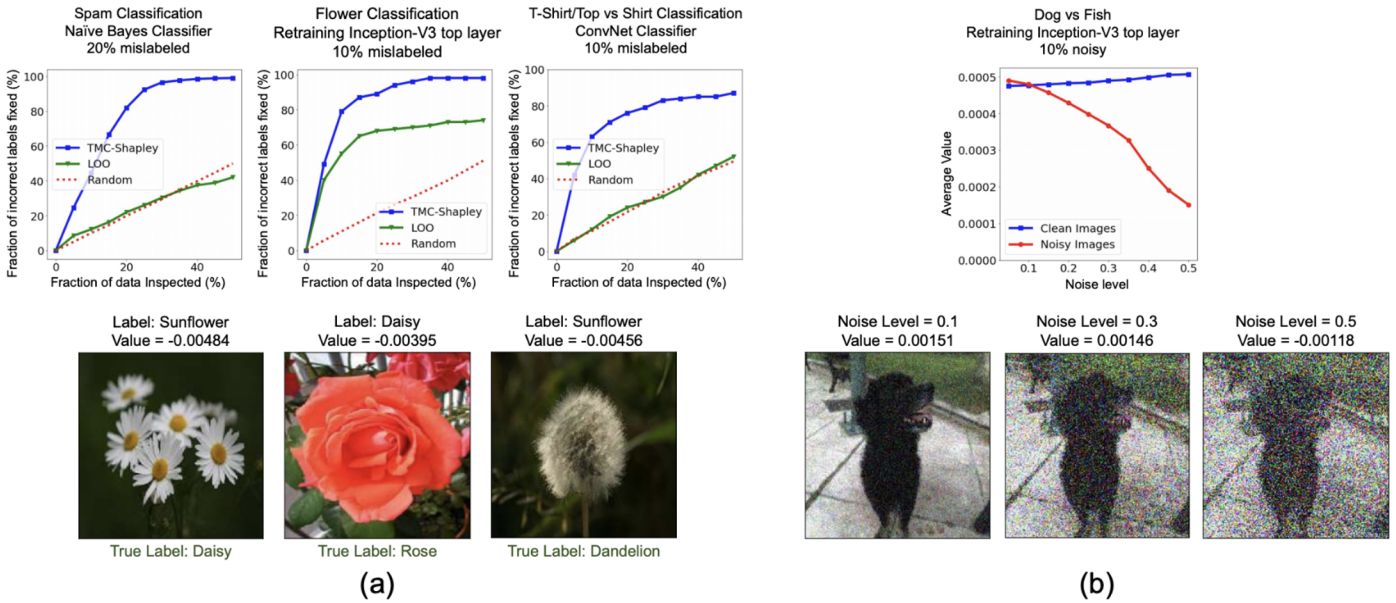


Figure 2. (a) We inspect train data points from the least valuable to the most valuable and identify the mislabeled examples. As it is shown, by using Shapley value we need the least number of inspections for detecting mislabeled data. While leave-one-out works reasonably well on the Logistic Regression model, it's performance on the two other models is similar to random inspection. (b) We add white noise to 10% of train data images. As the noise level increases, the average value of noisy images compared to clean images decreases. Each point on the plots is the average result for 10 repeats of the experiments where each time a different subset of train data is corrupted.

Experimental Analysis - 2

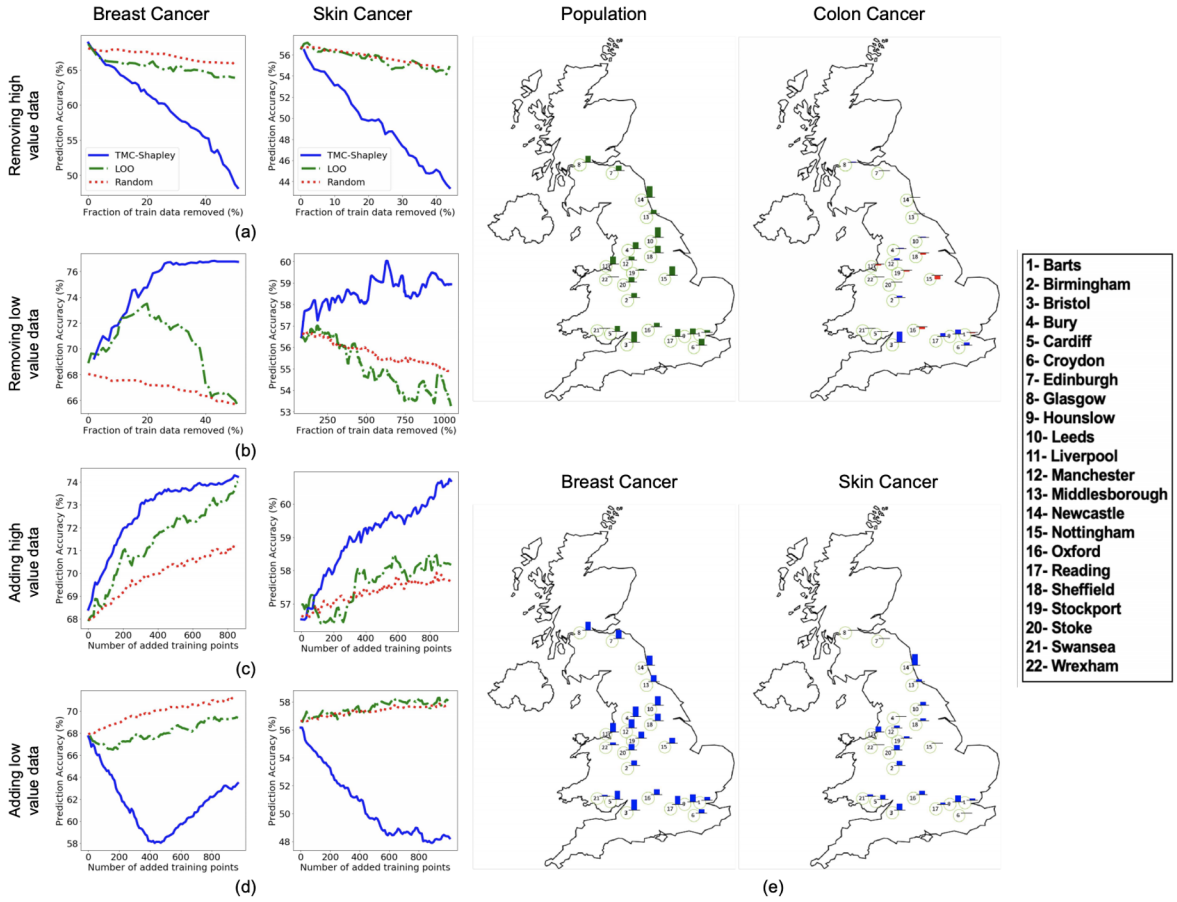
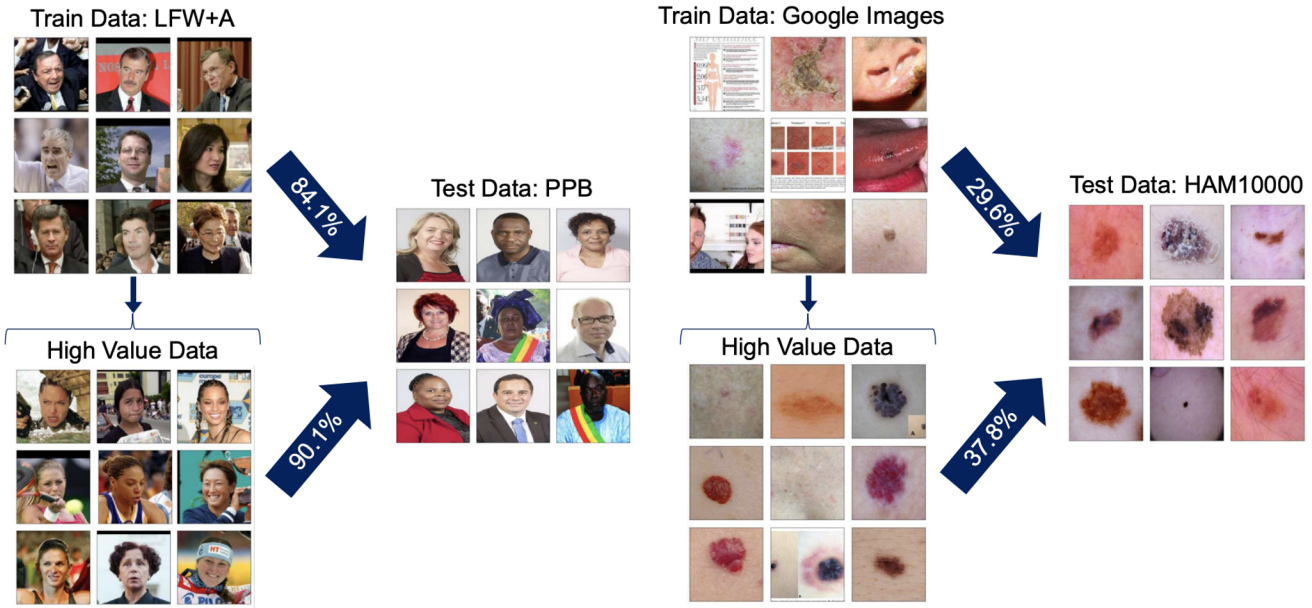


Figure 3. Patient Value for Disease Prediction For breast and skin cancer prediction tasks, we calculate the value of patient in the training data. (a) We remove the most valuable data from the train set and track the performance degradation. (b) Removing low value training data improves the predictor performance. (c) Acquiring new patients similar to high value training points improves performance more than adding patients randomly.(d) Acquiring new patients who are similar to low value points does not help. (e) Map of values of different centers across UK .

Experimental Analysis - 3

Source to Target	Prediction Task	Trained Model	Original Performance (%)	Adapted Performance (%)
Google to HAM1000	Skin Lesion Classification	Retraining Inception-V3 top layer	29.6	37.8
CSU to PP	Disease Coding	Retraining DeepTag top layer	87.5	90.1
LFW+ to PPB	Gender Detection	Retraining Inception-V3 top layer	84.1	91.5
MNIST to UPS	Digit Recognition	Multinomial Logistic Regression	30.8	39.1
Email to SMS	Spam Detection	Niave Bayes	68.4	86.4

(a)



(b)

Figure 4. Data shapley value for domain adaptation Adapting to a new data set. Available training data is not always completely similar to the test data. By valuating the training set data points, we can first, remove points with negative value and then, emphasize the importance of valuable points by assigning more weight during training.

Conclusion and Future Work

- Proposed a new technique for assigning values to data
- Can be used for increasing performance by rooting out the low Shapley value points