Computers & Operations Research 159 (2023) 106305

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

Variance reduced Shapley value estimation for trustworthy data valuation

Mengmeng Wu [a], Ruoxi Jia [b], Changle Lin [c], Wei Huang [d], Xiangyu Chang [a,*]

[a] School of Management, Xi'an Jiaotong University, Xi'an 710049, China
[b] Virginia Tech, Blacksburg, VA 24061, USA
[c] Institute for Interdisciplinary Information Core Technology, Xi'an 710065, China
[d] College of Business, Southern University of Science and Technology, Shenzhen 518055, China

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Machine learning
Data valuation
Stratified sampling
Variance reduction
Data Shapley

ABSTRACT

Data valuation, especially quantifying data value in algorithmic prediction and decision-making, is a fundamental problem in data trading scenarios. The most widely used method is to define the data Shapley and approximate it by means of the permutation sampling algorithm. To make up for the large estimation variance of the permutation sampling that hinders the development of the data marketplace, we propose a more robust data valuation method using stratified sampling, named variance reduced data Shapley (VRDS for short). We theoretically show how to stratify, how many samples are taken at each stratum, and the sample complexity analysis of VRDS. Finally, the effectiveness of VRDS is illustrated in different types of datasets and data removal applications.

1. Introduction

The emerging big data in all walks of life has become the driving force of technological and economic development (Ghorbani and Zou, 2019; Huang et al., 2021). Various sectors such as finance and healthcare increasingly rely on individuals' data for predictions, decision-making, and generating business value, which promotes extensive data transactions (Barua et al., 2012). One of the most critical problems in data trading scenarios is data valuation. We consider data trading scenarios in data markets based on machine learning models, such as DATABRIGHT (Dao et al., 2018) and Sterling (Hynes et al., 2018). The data value in this scenario is largely determined by its contribution to a specific machine learning model. We focus on data valuation in supervised learning, which is one of the main pillars of machine learning. The core challenge is how to fairly evaluate the contribution of each data in the training set to the learning algorithm for a particular performance metric.

A natural way to handle the aforementioned issue is to treat each data as a player in a cooperative game. Then, the value of each player can be assessed through utility functions from a game-theoretic perspective (Jia et al., 2019b). The Shapley value (Shapley, 2016), a solution concept that egalitarian distributes both gains and costs to several participants in a coalition, has been generalized to evaluate the contribution of each datum in supervised learning (Ghorbani and Zou, 2019; Jia et al., 2019a,b; Kwon and Zou, 2022; Tang et al., 2021). The Shapley value of a player is the average of its marginal contribution to all alliances that do not include itself. Since 2^{n-1} coalitions exclude it (assuming the number of players is n), the computational complexity of Shapley value is exponential. Indeed, the exact computation of Shapley value is NP-hard in general (Deng and Papadimitriou, 1994). How to effectively approximate the Shapley value of each data in supervised learning is the key to its application. The most widely used algorithm tackles this problem is the permutation sampling algorithm (also called Monte Carlo sampling) (Ghorbani and Zou, 2019; Jia et al., 2019b; Tang et al., 2021; Štrumbelj and Kononenko, 2014; Castro et al., 2009; Strumbelj and Kononenko, 2010; Cohen et al., 2007). One first samples a random permutation of training data, then scans one by one from the first element to the last element in the permutation, calculates the marginal contribution of each element to the set of elements in front of it, and finally repeats the same procedure over multiple permutations and takes the average of all their marginal contributions as the approximation of Shapley values. We refer to Shapley value for data valuation as data Shapley.

The permutation sampling gives an unbiased estimate of the data Shapley. However, it does not consider the impact of the cardinality of the training set on the model performance in machine learning. When we sample different permutations, the cardinality of the training set used to calculate the marginal contribution of each data may be different, resulting in relatively significant variances (see Fig. 1 for detailed discussions). However, the reproducibility of data valuation

* Corresponding author.
E-mail address: xiangyuchang@xjtu.edu.cn (X. Chang).

https://doi.org/10.1016/j.cor.2023.106305
Received 23 October 2022; Received in revised form 19 May 2023; Accepted 6 June 2023
Available online 21 June 2023

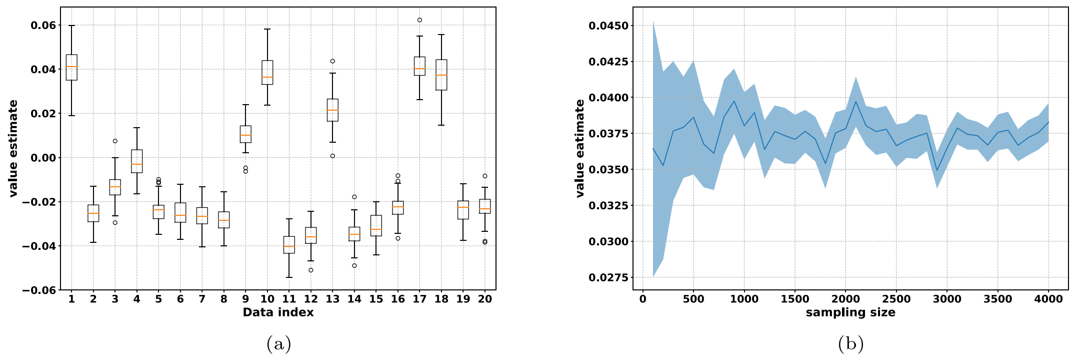


Fig. 1. Estimation value of 20 samples in the FashionMNIST dataset using Algorithm 1, in which the learning algorithm is the Naive Bayes (NB), and the utility function is the test accuracy. (a) shows the box plot of data Shapley, with a permutation sampling number of 1000 and 50 repetitions of the calculation. (b) shows the variation of variance of one data point as the sampling number increases. The solid line represents the mean of the estimated values of the data, and the colored area corresponds to the 5th and 95th confidence interval around the mean.

is the key to building trust in data transactions, and minimizing the variance of estimation results is critical to data valuation.

Our research aims to figure out the deficiency mentioned above of permutation sampling to estimate data Shapley and provide a more robust data valuation approach that can reduce the variance of estimating data Shapley. Therefore, we propose the following research questions (RQs): *What methods can be adopted to reduce the variance of evaluating data Shapley? What are their theoretical advantages and disadvantages? How to efficiently implement the proposed method in practice?* To answer these RQs, we are inspired by the so-called *stratified sampling* to propose a novel data Shapley estimation method called *Variance Reduced Data Shapley* (VRDS). Stratified sampling commonly divides the target population into several types or layers according to its attribute characteristics and then randomly selects samples from the layers (Cochran, 1977; Lohr, 2021). Stratified sampling increases the commonality of elements in each layer through classification and stratification, and it is easy to extract representative survey samples and further reduce the estimate variance. Moreover, we establish the total variance of stratified sampling for estimating data Shapley. To minimize the total variance, the optimal proportion of the number of sampled permutations per layer is manifested. The result is that the number of samples in each layer is directly proportional to the standard deviation of the marginal contribution in the corresponding layer. Since the actual standard deviation of each layer is unknown, how to effectively approximate the variance is a problem that must be considered. We justify that the function of set cardinality can be used as an upper bound on the variance, and approximate the variance by it, and then obtain the number of samples of each layer. Finally, the sample complexity analysis has been conducted for VRDS to demonstrate the theoretical property compared with the permutation sampling.

The rest of this paper is organized as follows. We discuss related works and contributions in Section 2. Section 3 introduces the data Shapley in supervised learning and presents the permutation sampling algorithm for estimating data Shapley. We propose the VRDS, including a stratified sampling algorithm and sample complexity analysis in Section 4. The experimental study of VRDS is provided in Section 5. Section 6 summarizes the conclusions. All the technical proofs are presented in Appendix.

2. Related work and contribution

Data valuation or pricing has become a significant role in the digital economy. Thus, various data valuation schemes have been studied in the literature. As we all know, data product replication costs are meager, even close to zero in many scenarios, so the traditional pricing strategy can no longer be directly applied to data valuation (Pei, 2020). In recent years, many articles have emerged to study different fields of data valuation, including arbitrage-free valuation (Li et al., 2014; Chen et al., 2019), revenue maximization valuation (Chawla et al., 2019), fair and truthful valuation (Ghorbani and Zou, 2019; Jia et al., 2019b), and privacy-preserving valuation (Li and Raghunathan, 2014). This paper aims to study the fair and truthful valuation of data, that is, price each data fairly according to the impact of data on the model performance in supervised learning.

Shapley value was introduced in the classical game theory (Shapley, 2016). Recent studies have adopted the Shapley value to measure the importance of inputs to output in a specific system. For example, suppose the specific system can be presented as a deterministic function. In that case, studies examine the impact of input variables on outputs by exploring their contribution to the variance of output, referred to as variance-based sensitivity analysis (Owen, 2014; Song et al., 2016; Owen and Prieur, 2017; Iooss and Prieur, 2019; Broto et al., 2020; Bénard et al., 2022; Herin et al., 2022; Da Veiga et al., 2021). Several approximate methods have been proposed in this field, such as Monte Carlo algorithm (Song et al., 2016), meta-model (Iooss and Prieur, 2019), double Monte Carlo (Broto et al., 2020), and Pick and Freeze estimators (Broto et al., 2020). In contrast, there are significant differences between these papers and our work in data valuation. Firstly, they assess the impact of input variables or features on output, which differs from our consideration of the impact of samples on output. Secondly, in variance-based sensitivity analysis, the functional relationship between input and output is assumed to be predetermined, but in the data valuation we study, the function is learned through training on samples. Lastly, most of them aim to measure the importance of variables through variance decomposition, which is different from our goal of reducing the total variance of data value estimation.

Another research line assumes the system is the supervised learning model. This case studies the significance of features or data in supervised learning settings by analyzing their impact on the performance of the learning model (Ghorbani and Zou, 2019; Jia et al., 2019b,a; Tang et al., 2021; Štrumbelj and Kononenko, 2014; Strumbelj and Kononenko, 2010; Cohen et al., 2007). To mitigate the substantial computational expense of Shapley value, several approximate algorithms have been developed in these studies. These approximation approaches include the permutation sampling algorithm (Ghorbani and Zou, 2019; Jia et al., 2019b), gradient Shapley algorithm (Ghorbani and Zou, 2019), group testing-based algorithm (Jia et al., 2019b). Although several approximation algorithms exist, the permutation sampling algorithm is the most frequently used and has been applied to data valuation (Ghorbani and Zou, 2019; Jia et al., 2019b,a; Kwon and Zou, 2022; Tang et al., 2021), feature attribution (Strumbelj and Kononenko, 2010; Cohen et al., 2007), cooperating game evaluation (Štrumbelj and Kononenko, 2014; Castro et al., 2009). Therefore, the permutation sampling algorithm has been confirmed as "the most useful baseline" for approximately computing Shapley value. Nonetheless, this algorithm suffers from significant variance resulting from random sampling, which limits its wide application (see Fig. 1 for discussions). Our study concentrates on data valuation in supervised machine learning and seeks to address the limitations of the permutation sampling algorithm by utilizing a different method, namely stratified sampling.

The stratified sampling has been considered in the cooperative game theory. Maleki et al. (2013) use the stratified sampling to improve the bound of the approximation error. That is, they first scale the difference between Shapley and its estimated value to obtain the expression related to the sampling number of each layer. By minimizing this difference, they decide the sampling number of each layer and give the specific implementation method of stratified sampling. Castro et al. (2017) employ stratified sampling to estimate Shapley value, by reducing the variance of the estimations obtained by stratified sampling giving the formula of sampling number in every stratum. They consider that one of the sources of variance for each marginal contribution is the player and the order in which that player arrives, so they take two sources of variation when stratifying. Burgess and Chapman (2021) derive a concentration inequality that is tailored to stratified Shapley value estimation using sample variance information. Based on this error bound, they propose an online process of sequentially choosing samples from the strata in order to minimize the estimate error. Our research focuses more on how to apply the Shapley value to the data pricing market. In the data market scenario, the large variance brought by algorithms for estimating data Shapley is unacceptable. Therefore, we consider using the stratified sampling method, but unlike (Maleki et al., 2013), we aim at the minimum variance of estimators from the algorithmic randomness. The algorithm in Castro et al. (2017) applies to situations where the variability of the marginal contributions depends greatly on each player's arrival position and is also not suitable for our scenario. While the algorithm in Burgess and Chapman (2021) selects samples online according to the conclusion of the central inequality to reduce the estimation error, which increases the unnecessary calculation cost for the data markets and does not consider the variance of the estimated value. Therefore, we propose a trustworthy data pricing approach specifically for data markets.

We summarize the contributions of this paper as follows:

- Proposing a trustworthy data Shapley estimation method, VRDS, based on the stratified sampling. The optimal number of samples in each layer is determined by minimizing the variance of VRDS. We justify that the variance of VRDS is less than or equal to the variance estimated by the permutation sampling algorithm.
- Implementing a stratified sampling algorithm for estimating VRDS. We theoretically provide the sample complexity analysis of the algorithm. The theoretical result indicates that the sample number of proposed VRDS based on the stratified sampling has the same order with the permutation sampling up to a log factor for achieving an (ϵ, δ)-approximation.
- Designing expensive experimental studies on rich data sets to illustrate the effectiveness of VRDS. It is found that it can significantly reduce variance and effectively identify data quality.

3. Preliminaries

In this section, we review Shapley value's concept and basic properties, based on which we set the framework for data valuation. We then introduce data Shapley for supervised learning and discuss the baseline permutation sampling algorithm to approximate the data Shapley.

3.1. Shapley value

The Shapley value is a classic concept in cooperative game theory (Shapley, 2016). A cooperative game is defined by a tuple $\langle N, U\rangle$, where $N = \{1, 2, \ldots, n\}$ which denotes the set of all players, and $U : 2^N \rightarrow \mathbb{R}$ is a map that assigns to each coalition $S \subseteq N$ a real number $U(S)$ such that $U(\emptyset) = 0$. It represents the utility of the collaboration of the members in S. The goal is to distribute the total income $U(N)$ among players according to each player's contribution to the cooperation. Shapley value of player i is defined as the average marginal contribution of i to all coalition that excludes i

$$\phi_i(U) = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}, |S|=s} \frac{1}{\binom{n-1}{s}} [U(S \cup \{i\}) - U(S)], \quad i = 1, \ldots, n, \tag{1}$$

where $U(S \cup \{i\}) - U(S)$ is the marginal contribution of player i with respect to S.

It has been justified that Shapley value is the unique solution for cooperative games satisfying the following axioms (Shapley, 2016), which prompts researchers to use it for data valuation (Ghorbani and Zou, 2019; Jia et al., 2019b).

Axiom 1 (*Efficiency Axiom*). $\sum_{i \in N} \phi_i(U) = U(N)$.

Axiom 2 (*Symmetry Axiom*). *For players i and j, if $U(S \cup \{i\}) = U(S \cup \{j\})$ holds for all S, where $S \subseteq N$ and $i, j \notin S$, then $\phi_i(U) = \phi_j(U)$.*

Axiom 3 (*Dummy Axiom*). *If $U(S \cup \{i\}) = U(S)$ holds for all S, where $S \subset N$ and $i \notin S$, then $\phi_i(U) = 0$.*

Axiom 4 (*Additivity Axiom*). *For any pair of games (N, U) and (N, V), if $(U + V)(S) = U(S) + V(S)$ holds for all S, then $\phi(U + V) = \phi(U) + \phi(V)$.*

The efficiency axiom states that players would expect to distribute all utility of their coalition. The symmetry axiom means that if two players have the same marginal contribution to all coalitions excluding them, they should have the same value. The dummy axiom requires value assignment should be sensitive to the player's contribution to all coalitions. A null player's marginal contribution is zero to any coalition that does not include itself. The additivity axiom can decompose a given utility function into an arbitrary sum of utility functions, and the value can be calculated respectively. Shapley value satisfies the above four axioms, so it is considered as an egalitarian distribution of cooperative games. This prompts scholars to regard the process of data co-training the model as a game, regard each data as a player, measure each data's contribution to the model with Shapley value, and conduct data pricing based on it.

3.2. Data Shapley for supervised learning

Consider a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ containing n samples, where x_i describes features of the ith instance and y_i corresponds to its label. We focus on the value of data in D. Given a learning algorithm $\mathcal{A}$ and a performance metric U to evaluate how well the model performs, we train it on a subset of D denoted by S and use a test set to compute the corresponding utility function $U(S)$. If it is a regression model, the performance metrics can be MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 (R Squared), and so on. If it is a classification model, the performance indicators can include accuracy, F1 score, etc. For example, if we consider a classification problem and select predictive accuracy as the utility function U, then $U(S) = \frac{1}{|T|} \sum_{j \in T} 1(y_j = \hat{g}_S(x_j))$, where $\hat{g}_S$ is the model trained on S and $|T|$ is the sample size of T. We aim to evaluate which data points significantly contribute to the corresponding evaluation metric U. We can formulate the data valuation problem as a cooperative game by treating the data as players and the evaluation metric as a utility function. Then the Shapley value could be employed to evaluate the value of data. The Shapley value of each data in supervised learning is referred to *data Shapley*.

The main challenge in adopting data Shapley is its computational cost. The computational complexity of evaluating the exact data Shapley using Eq. (1) requires $\mathcal{O}(2^n)$, since it involves computing marginal contributions of all points to all sets. In addition, for most algorithms, we need to train the machine learning model twice every time to calculate every marginal contribution, which is computationally expensive.

Many works of literature (Ghorbani and Zou, 2019; Jia et al., 2019b) have studied how to approximate data Shapley effectively. Up to now, the permutation sampling algorithm is a widely used baseline algorithm (see Algorithm 1). Recall that Shapley value has an equivalent form (see Ghorbani and Zou, 2019) as

$$\phi_i(U) = \mathbb{E}_{O\sim\Pi(N)}(U(P_i^O\cup\{i\})-U(P_i^O)) = \frac{1}{n!}\sum_{O\in\Pi(N)}(U(P_i^O\cup\{i\})-U(P_i^O)), \tag{2}$$

where $\Pi(N)$ is all permutations of N, and P_i^O represents the set of predecessors of i in the permutation O. That is, if $O = \{i_1, i_2, \ldots, i_k, i, i_{k+1}, \ldots, i_{n-1}\}$, $P_i^O = \{i_1, i_2, \ldots, i_k\}$. $P_i^O = \emptyset$ if i is the first element. According to the above formulation, we can consider that each permutation is uniformly sampled from $n!$ permutations with the probability $1/n!$. Intuitively, imagine all data are to be collected in a random order, and the marginal contribution of every data is to those already collected data. If we average these marginal contributions over all possible data orders, we obtain ϕ_i. Thus, we can regard ϕ_i as the expectation of marginal contributions and estimate it by the sample mean. The estimator $\hat{\phi}_i$ of ϕ_i is

$$\hat{\phi}_i(U) = \frac{1}{m}\sum_{O\in M}(U(P_i^O\cup\{i\})-U(P_i^O)), \tag{3}$$

where M represents permutation samples and $|M| = m$.

Algorithm 1 Permutation Sampling Algorithm for Data Shapley

Input: Training data $D = \{(x_i, y_i)\}_{i=1}^n$, learning algorithm $\mathcal{A}$, performance score U, sampling number m.
Output: Data Shapley of training data: $\phi_1, \ldots, \phi_n$.

Initialize $\phi_i := 0$ for $i = 1, \ldots, n$;
for $t = 0$ to m **do**
  O^t: random permutation of training data;
  $u_0^t := U(\emptyset, \mathcal{A})$;
  for $j = 1$ to n **do**
    $u_j^t := U(P_j^{O^t}, \mathcal{A})$;
    $\phi_{O^t[j]} := \frac{t-1}{t}\phi_{O^{t-1}[j]} + \frac{1}{t}(u_j^t - u_{j-1}^t)$;
  end for
end for

To evaluate the sufficient sampling number to achieve a specific error level, we denote the following (ϵ,δ)-approximation of data Shapley.

Definition 1. $\hat{\phi}$ is an (ϵ,δ)-approximation to ϕ if $\mathbb{P}[|\hat{\phi}-\phi| \le \epsilon] \ge 1-\delta$.

Lemma 1 provides a lower bound on the sampling number m in the permutation sampling algorithm to achieve an (ϵ,δ)-approximation.

Lemma 1. *Algorithm 1 returns an (ϵ,δ)- approximation to data Shapley of one data if the sampling number of permutations m satisfies $m \ge r^2\log(2/\delta)/2\epsilon^2$, where r is the range of the data's marginal contributions. In particular, if the utility function is the prediction accuracy, setting $m \ge \log(2/\delta)/2\epsilon^2$ is sufficient to achieve the (ϵ,δ)-approximation.*

In machine learning, the range of utility functions is usually determined. For example, if we choose the utility function as the prediction accuracy, then $0 \le U \le 1$. Therefore, the range of the marginal contributions of all data is $-1 \le r \le 1$, which makes the result of Lemma 1 relatively simple and intuitive. It is worth noting that this lemma has been proved by Maleki et al. (2013). We refine the analysis for comparison purposes in subsequent sample complexity analyses of proposed VRDS.

Although the permutation sampling algorithm is easy to operate, sometimes the variance of the data Shapley estimator is too large to be accepted for the data marketplace. For instance, Fig. 1(a) shows the box plot of data Shapley of 20 data from the FashionMNIST dataset (Xiao et al., 2017), in which the learning algorithm is the Naive Bayes (NB), the utility function is the test accuracy, the sampling number of permutation is 1000, and the calculation is repeated 50 times. Fig. 1(b) is the variation of the variance of one data point as the sampling number increases. The solid line is the mean of the estimated values of one data repeating 50 times, and the colored area is the 5th and 95th confidence interval around the mean. These two figures show that even with a large sampling number, the variance is too large compared with the value, which may lead to divergence and failure of transactions in practical application. Inspired by this, we propose the variance-reduced data Shapley in the following section.

4. Variance reduced data Shapley

This section utilizes stratified sampling to replace the random sampling in robustly estimating data Shapley.

4.1. Stratified sampling for data Shapley

The stratified sampling approach (Cochran, 1977) divides the population into several layers according to a specific characteristic and then randomly samples in each layer to form samples (Lohr, 2000).

For data Shapley, the utility function is usually an evaluation metric such as predictive accuracy, F1, etc. The values of evaluation metrics are directly affected by the training sample size. Therefore, for every data, we consider dividing the coalition sets according to their size when calculating the marginal contribution. Recall $\phi_i(U)$ that can be calculated as

$$\phi_i(U) = \frac{1}{n}\sum_{k=0}^{n-1}\sum_{S\subseteq N\setminus\{i\},|S|=k}\frac{1}{\binom{n-1}{k}}(U(S\cup\{i\})-U(S)). \tag{4}$$

In words, by grouping the coalitions which do not contain the data i, based on their sizes, we have n strata $S^0, S^1, \ldots, S^{n-1}$ such that $S^k = \{S \subseteq N\setminus\{i\}, |S| = k\}$ contains all the coalitions with size k. For example, suppose we have 5 samples and the index set is $\{1,2,3,4,5\}$. Now we want to calculate the Shapley value of the first datum. We need to know the marginal contribution of the first datum for all data sets that do not contain it. To accomplish this, we divide these sets into five categories based on their cardinality, list them separately, and then randomly sample m_k sets from the kth stratum to calculate, refer to Fig. 2. We suppose the dependency on U when the utility is self-evident and use ϕ_i to represent the value allocated to data i. Let $\phi_{i,k}$ denote the expected marginal contribution of the data i within stratum S^k, then we denote

$$\phi_{i,k} = \frac{1}{\binom{n-1}{k}}\sum_{S\subseteq N\setminus\{i\},|S|=k}(U(S\cup\{i\})-U(S)), \tag{5}$$

and it is obvious that data Shapley of data i can be calculated as follows

$$\phi_i = \frac{1}{n}\sum_{k=0}^{n-1}\phi_{i,k}. \tag{6}$$

Suppose that the random variable $X_i(S) = U(S\cup\{i\}) - U(S)$ represents the marginal contribution of data i to S, and $\phi_{i,k}$ is its expected marginal contribution to all sets that do not contain i and whose size is k. It is natural to use the sample mean to estimate its expectation to reduce computing costs. By denoting m_k the number of samples taken from S^k, we obtain the estimate of ϕ_i as

$$\hat{\phi}_i = \frac{1}{n}\sum_{k=0}^{n-1}\hat{\phi}_{i,k} = \frac{1}{n}\sum_{k=0}^{n-1}\frac{1}{m_k}\sum_{j=1}^{m_k}X_{i,k,j}, \tag{7}$$

where $X_{i,k,j}$ represents the marginal contribution of the ith point to the jth sample in stratum S^k. To minimize the variance of $\hat{\phi}_i$, a critical problem is determining the number of samples m_k in each stratum. Theorem 1 provides a feasible solution.

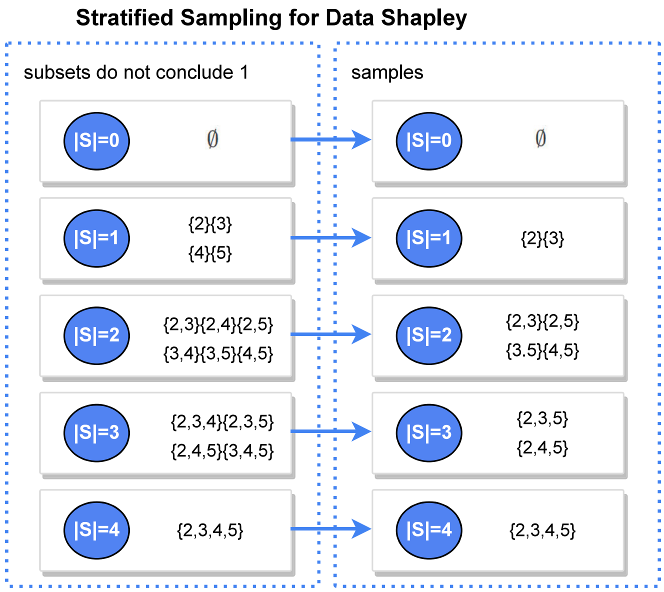


Fig. 2. Example of the stratified sampling.

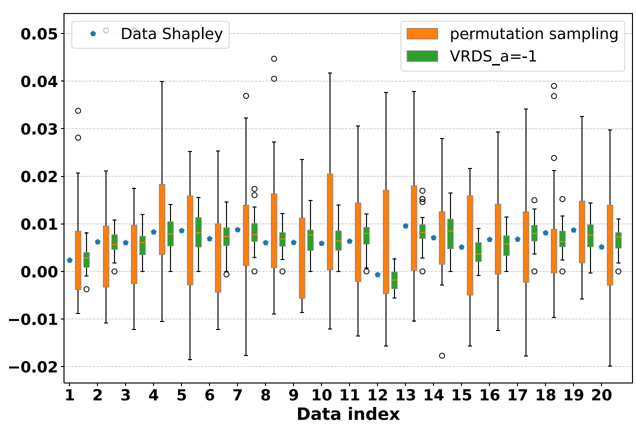


Fig. 3. The data Shapley values of 20 data in the FashionMNIST dataset, and their estimates. The learning algorithm is KNN and the utility function is test accuracy. Permutation sampling and $VRDS(a = -1)$ are used, respectively. The experiment is repeated 30 times to estimate values.

Theorem 1. *Suppose that using a training set D which contains n i.i.d samples to train a machine learning model and performs on a test set T. A utility function $U : 2^N \to \mathbb{R}$ satisfies the above four axioms. We utilize the proposed stratified sampling algorithm to estimate the data Shapley (Algorithm 2) with the total number of samples m. To minimize the variance of the data Shapley estimator, the optimal number of samples in the kth stratum is given by*

$$m^*_{i,k} = m\frac{\sigma_{i,k}}{\sum_{j=0}^{n-1}\sigma_{i,j}}, \quad k = 0, \dots, n-1, \tag{8}$$

where $\sigma_{i,k}$ denotes the standard deviation of $\hat{\phi}_{i,k}$. In addition, it can be proved that the variance obtained by using stratified sampling is no greater than that resulting from permutation sampling.

Theorem 1 shows that the variance obtained by the stratified sampling method is less than or equal to that obtained by the permutation sampling method. Moreover, it reflects that when the total number of samples m is fixed, the number of samples in each layer should be proportional to its standard deviation. In practical application, we do not know the variance of each layer, so it is necessary to provide an alternative value of variance to determine the sample size. Indeed, we can justify that the variance of each layer is bounded by the range of the utility function in that layer. It is worth noting that the variance does not increase in machine learning as the number of training samples increases. Therefore, it is reasonable to assume that the utility range does not increase for the number of layers k (the size of training set of samples in each layer increases for k). Taking full advantage of the above property, Theorem 2 provides a method to allocate the number of samples when the variance is unknown.

Theorem 2. *With the same assumptions of Theorem 1, we further denote that $r_{i,k}, i = 1, \dots, n, k = 0, \dots, n-1$ is the range of set $\{U(S \cup \{i\}) - U(S), \text{where } |S| = k\}$. Then there exists a non-increasing function of $f(k)$ that satisfies*

$$\sigma^2_{i,k} \le r^2_{i,k}/4 \le \frac{d_i^2 f(k)^2}{16}, \tag{9}$$

where d_i is a constant related to the ith sample. Then, using the upper bound of $\sigma^2_{i,k}$ to replace its functionality in (8), we can obtain the approximate m_k as

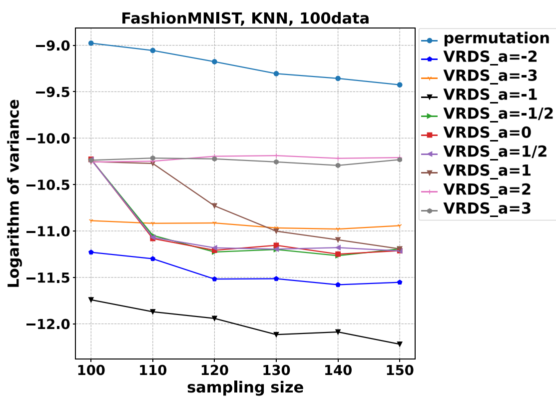
$$\tilde{m}_k = m\frac{f(k)}{\sum_{j=0}^{n-1} f(j)}, \quad k = 0, \dots, n-1.$$

Theorem 2 proves that the sample size of each stratum can be set as proportional to some non-increasing functions of k. Next, we analyze several specific forms of function f.

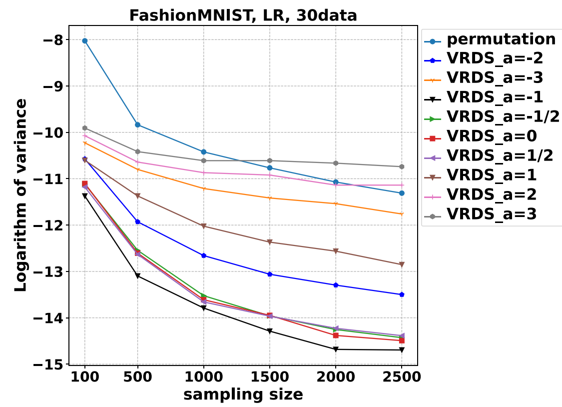
- $f(k) = c$, where c is a constant, then
$$\tilde{m}_k = m\frac{c}{\sum_{j=0}^{n-1} c} = \frac{m}{n}.$$
This means that the number of samples is equally distributed to each layer.
- $f(k) = \frac{1}{k+1}$, then
$$\tilde{m}_k = m\frac{\frac{1}{k+1}}{\sum_{j=0}^{n-1}\frac{1}{j+1}}.$$
Contrary to the above, assuming that $f(k)$ is decreasing with respect to k, because the number of samples per layer is also decreasing with respect to the set cardinality.
- $f(k) = (k+1)^a, a < 0, a \neq -1$, where the absolute value of a reflects the rate at which f changes with k. In different data sets, different algorithms are used to calculate different amounts of data Shapley. Thus, a can be considered as a tuning parameter that can influence the performance of estimating data Shapley. In machine learning applications, the utility function of a set S is often defined as the loss of the model trained over S for predicting a test point z, i.e., $U(S) = l(\mathcal{A}(S), z)$, where $\mathcal{A}$ represents the underlying learning algorithm that takes in a dataset and outputs a model. Under this utility function definition, the range of $U(S \cup \{i\}) - U(S)$—the marginal contribution of any data point i to a subset of size k—can be upper bounded by uniform stability of $\mathcal{A}$, defined by $\max_{S \in \mathcal{Z}^k} \max_z \max_i |l(\mathcal{A}(S \cup \{i\}), z) - l(\mathcal{A}(S), z)|$. Prior work (Bousquet and Elisseeff, 2002; Hardt et al., 2016) have shown that the upper bound of uniform stability of many common learning algorithm at size k is $\mathcal{O}(\frac{1}{k+1})$. These theoretical results shed light on our empirical observation that the best choice of a is usually -1, giving rise to the least variance in data Shapley estimation compared to the other possible choices. Therefore, we provide that a suggested interval of a is $[-1, -1/2]$. Detailed pieces of evidence will be presented in the experimental section.

Taking into account that sample size should be an integer, we can set the value of $m_k = \min\{1, \lfloor\tilde{m}_k\rfloor\}$. However, this implies that additional samples may be left unused as $\sum_{k=0}^{n-1} m_k$ may be lower than m. In this case, we sequentially increase the value of m_k from $k = 0$ to $n-1$ until the sum exceed m.

So far, we have handled how to use the stratified sampling to reduce the variance of estimated data Shapley. Next, we will give the specific algorithm for the real implementation in Algorithm 2.



(a)



(b)

Fig. 4. The variation of variance with the number of samples when using the permutation sampling algorithm and the VRDS algorithm where the parameter a takes value ranging from −3 to 3. The dataset is FashionMNIST. In Fig. (a), the KNN model is used, and the dataset size consists of 100 points. In Fig. (b), the LR model is used, and we consider 30 data points.

4.2. VRDS and its sample complexity analysis

Algorithm 2 presents the pseudo-code of the stratified sampling algorithm, which first calculates the sampling number of each stratum and then derives the value of each data by taking samples from every stratum. Note that we refer to the estimated data Shapley by Algorithm 2 as variance reduced data Shapley (VRDS for short).

Algorithm 2 Stratified Sampling Algorithm for VRDS

Input: Training data $D = \{(x_i, y_i)\}_{i=1}^n$, learning algorithm $\mathcal{A}$, performance score U, and sampling number m.
Output: Shapley value of training data: $\phi_1, \dots, \phi_n$

```
Initialize φ_i := 0 for i = 1,…,n and t = 0;
m_k := min{1, ⌊m f(k) / Σ_{j=0}^{n-1} f(j)⌋}, for k = 0,…,n−1;
while m − Σ m_k > 1 do
  m_t := m_t + 1; t := t + 1;
end while
for i ∈ {1,…,n} do
  for k ∈ {0,…,n−1} do
    l := 0, φ_{i,k} := 0;
    while l ≤ m_k do
      S := get a random coalition of {1,…,i−1,i+1,…,n} with size k;
      φ_{i,k} := φ_{i,k} + (U(S ∪ {i}, A) − U(S, A));
      l := l + 1;
    end while
    φ_{i,k} := φ_{i,k} / m_k;
    φ_i := φ_i + φ_{i,k} / n;
  end for
end for
```

In this section, we will present sample complexity analysis of VRDS to explain the sufficient sampling number in the stratified sampling algorithm to achieve an (ϵ, δ)-approximation.

Theorem 3. *Algorithm 2 gets an (ϵ, δ)-approximation of data Shapley of each data if the sampling number m satisfies*

$$m \geq \max\left(\frac{16 \log \frac{2}{\delta}}{17\epsilon^2 n^2} \sum_{k=0}^{n-1} \frac{1}{f(k)} \sum_{j=0}^{n-1} f(j), \frac{2 \log \frac{2}{\delta}}{\epsilon^2 n^2 (f(n-1))^2} (\sum_{j=0}^{n-1} f(j))^2 \right). \tag{10}$$

When $f(k) = \frac{1}{k+1}$, it is sufficient to have

$$m \geq \frac{2 \log \frac{2}{\delta}}{\epsilon^2} (\log n + 1)^2. \tag{11}$$

Remark 1. From Lemma 1, we know that the minimum number of samples achieving (ϵ, δ)- approximation in the permutation sampling algorithm is independent of n. For the stratified sampling algorithm, since the sampling number depends on stratum number n, according to Theorem 3, it requires the same order (if we choose the suggested tuning parameter $a = -1$) with the permutation sampling up to a log factor about n for (ϵ, δ)- approximation. The stratified sampling algorithm achieves a small variance when the sample size m does not increase significantly with the increase of n, which provides a theoretical foundation for its wide application.

5. Experiments

5.1. Experimental setting

5.1.1. Experimental objective

Our experiments have the following three purposes.

First, we show that the stratified sampling algorithm for VRDS can estimate the data Shapley, and the variance of the estimated value by VRDS is smaller than that by the permutation sampling algorithm. Fortunately, Jia et al. (2019a) propose an elaborate method to compute the exact data Shapley for K-Nearest Neighbor (KNN) algorithm. Therefore, we use KNN in the first experiment.

Second, the performance of VRDS is directly affected by the tuning parameter a, which determines the number of samples per stratum. In order to explore whether different data sets, sample sizes and algorithms affect the selection of a, we use 14 data sets and five commonly used algorithms and conduct experiments with data sizes from 20 to 2000 to obtain the suggested parameter selection scheme of a.

Finally, data quality is a critical indicator for data analysis, mining, and application. It is usually necessary to judge the data quality in the data preprocessing stage and remove the data with poor quality (Wang et al., 2021). Data Shapley can be considered as a data quality measurement. So one of the applications of data Shapley is removing data with poor quality based on data Shapley for improving the prediction performance. In this part, we construct a new criterion based on the estimated variance of data Shapley to remove bad quality data and compare it with the results of the permutation sampling algorithm.

Table 1
Data sets.

Data set	Model	Reference
FashionMNIST	LR, KNN	Xiao et al. (2017)
Iris	NB, KNN	Fisher (1936)
Digits	KNN, Tree	Alimoglu and Alpaydin (1996)
Breast Cancer	SVC, KNN	Mangasarian et al. (1995)
Spam classification	NB, LR	Email Spam Classification Dataset CSV| Kaggle
Creditcard	LR	Yeh and Lien (2009)
Vehicle	LR	Duarte and Hu (2004)
Apsfail	LR	ml
Phoneme	LR	Keel
Wind	LR	wind
Pol	LR	pol
Cpu	LR	cpu
Fraud	LR	Dal Pozzolo et al. (2015)
2Dplanes	LR	2dplanes

5.1.2. Datasets

We evaluate the performance of the stratified sampling algorithm for VRDS on image data and tabular data. Table 1 describes these data sets. These data sets are chosen to provide classification problems, with varying dimensionality, and a mixture of problem domains. We set data size ranges from 20 to 2000 to verify the algorithm's robustness.

5.1.3. Machine learning model and performance metric

To prove that VRDS is effective for various algorithms, we use Logistic Regression (LR), K-Nearest Neighbor algorithm (KNN), Naive Bayes (NB), Decision Tree (DT), and Support Vector Classification (SVC). To implement these algorithms, we directly employed the Python package "scikit-learn". To accurately calculate the data Shapley value, another KNN solver in Jia et al. (2019a) is used. Since the performance metric selection has no direct relationship with the model performance, we always use prediction accuracy as the performance evaluation metric of all algorithms.

We set the sampling number from 100 to 5000 to conduct comparative experiments to test the effect of our VRDS and compare the deviation and variance of the estimated values obtained by VRDS and permutation sampling. For each experiment, we randomly sampled fifty times for estimation to obtain the variance of the estimated value.

5.2. Experimental results

Variance Comparison of VRDS. We first compare the data Shapley to its estimates obtained by employing two algorithms: permutation sampling and VRDS($a = -1$). To compute the data Shapley for large datasets, we utilize KNN since (Jia et al., 2019a) provides an effective and accurate approach for calculating data Shapley using KNN. Taking test accuracy as the utility function, we calculated the data Shapley of 100 data points and their estimates using permutation sampling and VRDS($a = -1$). Fig. 3 displays the data Shapley and their estimates for 20 of these data points. Both permutation sampling and VRDS($a = -1$) provide estimates close to exact values, but estimates from VRDS($a = -1$) have much lower variance compared to those from permutation sampling.

Parameter Selection for VRDS. We compare the variance reduction effect of VRDS when parameter a takes different values, and different data sets, sample sizes, and algorithms are considered. We first make intensive attempts on the parameters with the FashionMNIST dataset. Fig. 4(a) illustrates the variances of data Shapley estimates for 100 data points computed using the VRDS method with different values of parameter a, with the KNN machine learning model. The horizontal axis denotes the number of samples, ranging from 100 to 150, while the vertical axis represents the logarithm of the corresponding variance. Fig. 4(b) presents a scenario where estimates of 30 data Shapley are calculated and the machine learning model is LR. All estimates are obtained through 30 repetitions. From the two figures, we can see that in all methods, when the number of samples is fixed, VRDS with

Table 2
A summary of the variance of data Shapley estimator on different datasets when parameter a takes different values in the VRDS algorithm, with LR as the base model. The data volume is 100, the sampling number is 150, and the unit of variance is 10^{-6}. The best result is highlighted in bold. All estimates are obtained through 30 repetitions.

Dataset	Permutation	$a = 0$	$a = -1/2$	$a = -1$	$a = -2$
Cpu	70.36	4.05	2.18	**1.97**	2.50
Pol	59.15	2.36	1.76	**1.73**	2.39
Vehicle	58.22	2.60	2.63	**2.37**	3.09
2dplanes	59.79	2.42	2.41	**2.15**	3.02
Creditcard	55.42	2.31	**2.27**	2.55	3.19
Apsfail	76.24	3.86	2.21	**1.77**	2.27
Phoneme	62.29	2.97	2.59	**2.53**	3.18
Fraud	78.83	4.92	2.42	**1.84**	2.76
Wind	72.05	4.22	2.21	**1.81**	2.39

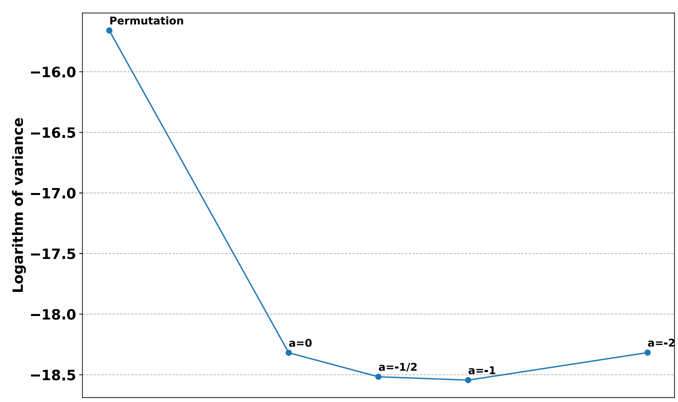


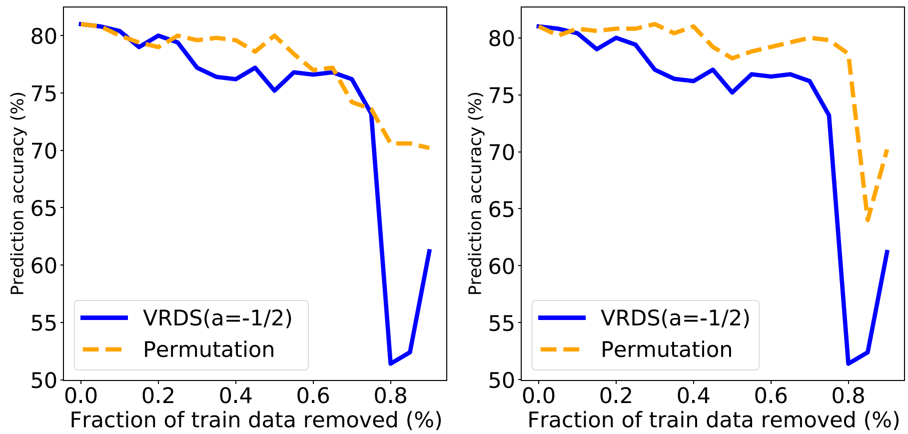
Fig. 5. Estimation of 2000 data Shapley in FashionMNIST data set, wherein the optimization algorithm is LR and the number of samples is 2200.

$a \in [-1, -1/2]$ has a minor variance (see more experiments in Appendix E). In addition, the variance decreases with the number of samples.

We use Table 2 to summarize the variance of the results calculated by different models on different data sets when the algorithm is LR and the number of samples is 150. It can be seen that the variance of the estimated value obtained by VRDS is obviously smaller than that obtained by the permutation sampling algorithm. Moreover, when $a = -1$ and $-\frac{1}{2}$, the VRDS method is better than when a takes other parameters.

We also examine whether VRDS is also effective when the data set is large. Fig. 5 shows data Shapley estimated by VRDS of 2000 points in the FashionMNIST data set. The algorithm is LR, and the number of samples is 2200. We can see that the variance of the results calculated by Permutation sampling is the largest, and the variance is the smallest when $a = -1$.

Data Group Removal. We evaluate the VRDS by comparing the performance on the data group removal task. The so-called data group



(a) (b)

Fig. 6. Data group removal experiment on the FashionMNIST data set. There are 20 groups and each group has 100 data. VRDS ($a = -\frac{1}{2}$) and permutation sampling algorithms are used to calculate the value of data groups. Fig. 6(a) shows the removal from the most valuable data group. Fig. 6(b) shows the removal from large to small according to the values of $estimated\ data\ Shapley - 100 \times variance$.

means that a data source contains multiple data. In fact, in a real data transaction scenario, a data provider usually provides a data set, and the data sets of providers jointly form a training set for a machine learning task. So in this experiment, we consider the value of the data group. The data group removal task is to remove groups of data to understand their impact on a model. A better data value estimation algorithm can better identify the importance of data. Therefore, when the data group with the highest (lowest) value estimation is removed, a better data value estimation method will result in faster (slower) performance degradation.

Fig. 6 shows the data group removal experiment on the FashionMNIST dataset. The data has been randomly partitioned into 20 groups, and each group consists of 100 data points. Notably, in practical applications, the dataset of each group generally derives from a single company.

VRDS with $a = -\frac{1}{2}$ and permutation sampling algorithms are used to calculate the value of data groups. Fig. 6(a) shows the removal from the most valuable data group. It can be seen that the performance of VRDS with $a = -\frac{1}{2}$ decreases faster than that of permutation sampling. Fig. 6(b) shows the removal according to the value of $estimated\ data\ Shapley - 100 \times variance$, the prediction accuracy of the VRDS with $a = -\frac{1}{2}$ algorithm decreases faster. This indicates that VRDS can identify important data groups more accurately by simultaneously combining the estimated data Shapley and its variance.

So far, we have demonstrated the effectiveness of VRDS in approximating data Shapley and obtained the suggested parameter a that minimizes the variance of the estimated value through various experiments on a large number of different data sets. Finally, the data group removal experiment shows that combining the data value and variance can quickly identify important data sets. To show more evidence of the results, we further present more experiments in Appendix E.

6. Conclusion

In this work, we propose a more robust data Shapley estimation method, VRDS, which is based on a stratified sampling technique and can obtain data Shapley estimates with minor variance. We obtain the optimal sampling number of each stratum in the stratified sampling and provide the sample complexity analysis. Experiments on various data sets show that the variance of estimates calculated by VRDS is always smaller than that calculated by the permutation sampling algorithm. Furthermore, we also get the suggested parameter interval that can obtain the estimation value with minimum variance on these data sets. Finally, the application of VRDS in data quality identification is also discussed. We find that simultaneously considering value and variance can recognize the most valuable data faster. For future work, we wish to continue applying the stratified sampling algorithm to other data Shapley estimation methods (Jia et al., 2019b,a; Kwon and Zou, 2022; Ghorbani et al., 2020) to further reduce the variance of data Shapley estimates to promote the development of the data marketplace.

CRediT authorship contribution statement

Mengmeng Wu: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Ruoxi Jia:** Methodology, Writing – review & editing, Visualization. **Changle Lin:** Data curation, Formal analysis. **Wei Huang:** Investigation, Visualization. **Xiangyu Chang:** Conceptualization, Methodology, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

We express our gratitude to Jiachen T. Wang, a Ph.D. student at Princeton University for his valuable comments on the manuscript and helpful discussions regarding its implementation. Additionally, Xiangyu Chang was supported in part by the National Natural Science Foundation for Outstanding Young Scholars of China under Grant 72122018 and in part by the Natural Science Foundation of Shaanxi Province under Grant 2021JC-01. Similarly, Ruoxi Jia would like to thank the support from NSF via the Grant OAC-2239622.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cor.2023.106305.

References

Alimoglu, F., Alpaydin, E., 1996. Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition. In: Proceedings of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium. Citeseer.

Barua, A., Mani, D., Mukherjee, R., 2012. Measuring the business impacts of effective data. p. 2012, Report accessed at http://www.sybase.com/files/White_Papers on Sep 15.

Bénard, C., Biau, G., Da Veiga, S., Scornet, E., 2022. SHAFF: Fast and consistent SHApley effect estimates via random forests. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 5563–5582.
Bousquet, O., Elisseeff, A., 2002. Stability and generalization. J. Mach. Learn. Res. 2, 499–526.
Broto, B., Bachoc, F., Depecker, M., 2020. Variance reduction for estimation of Shapley effects and adaptation to unknown input distribution. SIAM/ASA J. Uncertain. Quantif. 8 (2), 693–716.
Burgess, M.A., Chapman, A.C., 2021. Approximating the Shapley value using stratified empirical Bernstein sampling. In: International Joint Conference on Artificial Intelligence. pp. 73–81.
Castro, J., Gómez, D., Molina, E., Tejada, J., 2017. Improving polynomial estimation of the Shapley value by stratified random sampling with optimum allocation. Comput. Oper. Res. 82, 180–188.
Castro, J., Gómez, D., Tejada, J., 2009. Polynomial calculation of the Shapley value based on sampling. Comput. Oper. Res. 36 (5), 1726–1730.
Chawla, S., Deep, S., Koutrisw, P., Teng, Y., 2019. Revenue maximization for query pricing. Proc. VLDB Endow. 13 (1), 1–14.
Chen, L., Koutris, P., Kumar, A., 2019. Towards model-based pricing for machine learning in a data marketplace. In: Proceedings of the 2019 International Conference on Management of Data. pp. 1535–1552.
Cochran, W.G., 1977. Sampling Techniques. John Wiley & Sons.
Cohen, S., Dror, G., Ruppin, E., 2007. Feature selection via coalitional game theory. Neural Comput. 19 (7), 1939–1961.
Da Veiga, S., Gamboa, F., Iooss, B., Prieur, C., 2021. Basics and Trends in Sensitivity Analysis: Theory and Practice in R. SIAM.
Dal Pozzolo, A., Caelen, O., Johnson, R.A., Bontempi, G., 2015. Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence. IEEE, pp. 159–166.
Dao, D., Alistarh, D., Musat, C., Zhang, C., 2018. Databright: Towards a global exchange for decentralized data ownership and trusted computation. arXiv preprint arXiv:1802.04780.
Deng, X., Papadimitriou, C.H., 1994. On the complexity of cooperative solution concepts. Math. Oper. Res. 19 (2), 257–266.
Duarte, M.F., Hu, Y.H., 2004. Vehicle classification in distributed sensor networks. J. Parallel Distrib. Comput. 64 (7), 826–838.
Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7 (2), 179–188.
Ghorbani, A., Kim, M., Zou, J., 2020. A distributional framework for data valuation. In: International Conference on Machine Learning. PMLR, pp. 3535–3544.
Ghorbani, A., Zou, J., 2019. Data shapley: Equitable valuation of data for machine learning. In: International Conference on Machine Learning. PMLR, pp. 2242–2251.
Hardt, M., Recht, B., Singer, Y., 2016. Train faster, generalize better: Stability of stochastic gradient descent. In: International Conference on Machine Learning. PMLR, pp. 1225–1234.
Herin, M., Idrissi, M.I., Chabridon, V., Iooss, B., 2022. Proportional marginal effects for global sensitivity analysis. arXiv preprint arXiv:2210.13065.
Huang, L., Dou, Y., Liu, Y., Wang, J., Chen, G., Zhang, X., Wang, R., 2021. Toward a research framework to conceptualize data as a factor of production: The data marketplace perspective. Fund. Res. 1 (5), 586–594.
Hynes, N., Dao, D., Yan, D., Cheng, R., Song, D., 2018. A demonstration of sterling: a privacy-preserving data marketplace. Proc. VLDB Endow. 11 (12), 2086–2089.
Iooss, B., Prieur, C., 2019. Shapley effects for sensitivity analysis with correlated inputs: comparisons with Sobol'indices, numerical estimation and applications. Int. J. Uncertain. Quantif. 9 (5).
Jia, R., Dao, D., Wang, B., Hubis, F.A., Gürel, N.M., Li, B., Zhang, C., Spanos, C.J., Song, D., 2019a. Efficient task-specific data valuation for nearest neighbor algorithms. Proc. VLDB Endow. 12 (11), 1610–1623.
Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Gürel, N.M., Li, B., Zhang, C., Song, D., Spanos, C.J., 2019b. Towards efficient data valuation based on the shapley value. In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1167–1176.
Kwon, Y., Zou, J., 2022. Beta Shapley: a unified and noise-reduced data valuation framework for machine learning. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 8780–8802.
Li, C., Li, D.Y., Miklau, G., Suciu, D., 2014. A theory of pricing private data. ACM Trans. Database Syst. 39 (4), 1–28.
Li, X.-B., Raghunathan, S., 2014. Pricing and disseminating customer data with privacy awareness. Decis. Support Syst. 59, 63–73.
Lohr, S.L., 2000. Sampling: Design and analysis. Technometrics 42 (2), 223–224.
Lohr, S.L., 2021. Sampling: Design and Analysis. Chapman and Hall/CRC.
Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., Rogers, A., 2013. Bounding the estimation error of sampling-based Shapley value approximation. arXiv preprint arXiv:1306.4265.
Mangasarian, O.L., Street, W.N., Wolberg, W.H., 1995. Breast cancer diagnosis and prognosis via linear programming. Oper. Res. 43 (4), 570–577.
Owen, A.B., 2014. Sobol'indices and Shapley value. SIAM/ASA J. Uncertain. Quantif. 2 (1), 245–251.
Owen, A.B., Prieur, C., 2017. On Shapley value for measuring importance of dependent inputs. SIAM/ASA J. Uncertain. Quantif. 5 (1), 986–1002.
Pei, J., 2020. A survey on data pricing: from economics to data science. IEEE Trans. Knowl. Data Eng. 34 (10), 4586–4608.
Shapley, L.S., 2016. 17. A value for n-person games. In: Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press, pp. 307–318.
Song, E., Nelson, B.L., Staum, J., 2016. Shapley effects for global sensitivity analysis: Theory and computation. SIAM/ASA J. Uncertain. Quantif. 4 (1), 1060–1083.
Strumbelj, E., Kononenko, I., 2010. An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. 11, 1–18.
Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41 (3), 647–665.
Tang, S., Ghorbani, A., Yamashita, R., Rehman, S., Dunnmon, J.A., Zou, J., Rubin, D.L., 2021. Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. Sci. Rep. 11 (1), 1–9.
Wang, T., Zeng, Y., Jin, M., Jia, R., 2021. A unified framework for task-driven data quality management. arXiv preprint arXiv:2106.05484.
Xiao, H., Rasul, K., Vollgraf, R., 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.
Yeh, I.-C., Lien, C.-h., 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst. Appl. 36 (2), 2473–2480.