Towards a Subset-Sale Mode for Neural Networks in a Data Marketplace

Anonymous Author(s)

ABSTRACT

Data analysis using neural networks (NN) is becoming dominant in both academic and industrial domains. Buyers are eager to acquire datasets through the data marketplace for their NN model training. Previous work on data market mechanisms only focused on the fullset-sale mode (FsM) that forces buyers to purchase the whole dataset without any awareness of the data utility in the downstream model training. Such limitation detracts from the engagement and enthusiasm of budget-conscious buyers, which prevents data vendors from maximizing their revenue.

In this paper, we propose a subset-sale mode (SsM) data market that allows buyers to purchase subsets of a dataset in a variety of forms. We first formally describe the desired properties of the SsM framework, focusing on allocating equitable data value and formulating a subset value-based pricing (SVBP) strategy. Due to SVBP is dependent on the subsets' value, we then propose a novel data valuation method, *ClusterRVSV*, and present the theoretical proof of how *ClusterRVSV* can set equitable value to subsets even under dishonest behavior, e.g., replication and outliers. Based on the proposed data valuation method, we formulate an optimization problem for the SsM data market with the goal of maximizing data vendors' revenue while satisfying the necessary pricing constraints. Finally, we perform extensive experiments to validate the advantages of *ClusterRVSV* on subsets' data valuation, and also verify that the SsM data market can provide a higher revenue to data vendors and a higher affordability to data buyers.

1 INTRODUCTION

Neural networks (NN) have been used extensively in various applications, e.g., social media, health care, personal assistants, marketing and sales [1–3]. Datasets play a crucial role in the development of neural networks. As data-driven models, the performance of neural networks is highly dependent on the quality and quantity of datasets. As a result, buyers are eager for data market platforms that allow them to easily purchase/subscribe high-quality data from multiple sources. With the explosive growth of data supplies and demands, data exchange platforms (e.g., Dawex [4], WorldQuant [5]) have emerged, allowing buyers and vendors to discover, distribute, commercialize and exchange data freely. Such data marketplaces facilitate data circulation while also discovering the potential data value. Through the data market, buyers can obtain rich structured (relational) datasets to better train their NN models. In practice, datasets in the data market can be prohibitively expensive due to the high costs involved in the data collection, integration, and cleaning process. Additionally, existing pricing schemes operate on a fullset-sale mode (**FsM**), forcing buyers to purchase the whole dataset without any awareness of the data utility in the downstream model training. Due to the high price and unknown performance, FsM detracts from the engagement and enthusiasm of budget-conscious buyers. Such inefficient data markets fail to maximize data vendors' revenue and facilitate the data circulation. Given that not all dimensions of a dataset are equally valuable for the model training [6], a subset-sale mode data market that enables buyers to subscribe to subsets in a variety of forms is strongly desired.

The Subset-sale Mode(SsM) Data Market Framework. In this paper, we propose a framework for a SsM data market that enables buyers to purchase partial datasets at a lower budget. Our key observation is that, rather than purchasing the whole dataset, buyers prefer to purchase a subset containing certain dimensions at a lower price but sufficient for their model performance. Subsets of a dataset can be created with the help of well-developed data partitioning and sampling methods [7]. Now, the critical issue for the SsM data market is to design a pricing strategy that is monotone with respect to the value of the subset data. We refer to this pricing strategy as subset value-based pricing (**SVBP**). A high-level framework of the SsM data market is demonstrated in Figure 1. First, the market platform partitions the original dataset into subsets of varying sizes. The data market then determines the value and pricing of each subset. It is reasonable to regard the market as a trustworthy third party for data valuation. In exchange for a reward, a data market platform promises to facilitate mutual satisfaction in order to finalize deals. Following the pricing, the data description and data valuation results are packaged as the metadata and made available to buyers. Buyers can find their desired dimensions based on the metadata, which are more valuable to their NN models. Finally, a deal may be initiated when a buyer's willingness to pay (**WTP**) price for a subset is less than its price from the data vendor. Several examples explain why SsM is necessary for the data market in the following.

Practical Case 1 John is a researcher who wants to train a NN model to predict the vehicle flow at an intersection in New York City. Fortunately, he finds the tabular data related to this intersection's traffic flow on the data market. This dataset's feature columns contain statistical data about pedestrians, bicycles, and vehicles. Only vehicle-related features are needed for John's task. The FsM data market forces John to pay a high price for the entire dataset, which exceeds his budget. In this case, a SsM data market would enable John to charge a reasonable fee for a subset that includes his desired features.

Practical Case 2 Sharon is a specialist who wants to train a NN model to classify lung cancer in smokers over 50 years old. She finds a dataset containing related data of all people over 50 in the data market. She needs specific rows that are "yes" on the "smoking or not" attribute. In this case, SsM can help Sharon buy specified rows at a reasonable price, and the data vendors are also profitable.

Desiderata and Challenges. Because the pricing function is related to data value (data utility) [8] and SsM necessitates more frequent data valuation than FsM, the most technical challenge in achieving SsM is an efficient data valuation method on subsets. As illustrated in the Figure 1, data valuation is a bridge between the

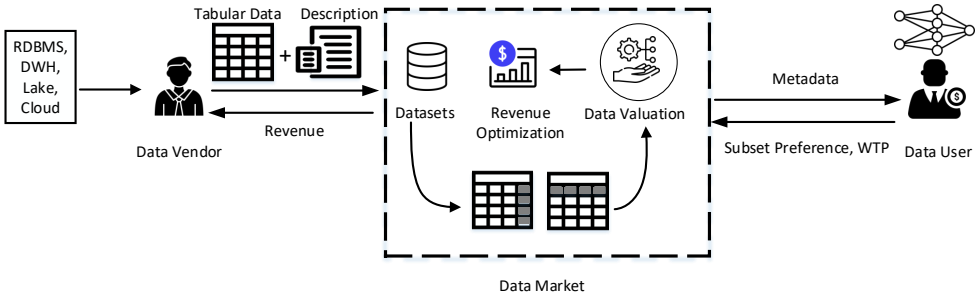

Figure 1: The framework of the subset-sale mode(SsM) market. The data market mainly includes three parties: (A) data vendors, also data owners, want to sell their datasets in exchange for certain revenue; (B) The data market is the platform to support dataset exchange; (C) data buyers, also owners of downstream models, want to pay for the desired subsets of a dataset to train their NN models.

buyers' awareness and the data vendors' revenue. To this end, the subset value-based pricing (SVBP) function in SsM data markets needs to meet the following properties. First, the SVBP aims to promote affordability and allow buyers with lower budgets to buy certain subsets. Thus, a subset should be allocated less value than its supersets. Second, to improve buyers' engagement and enthusiasm, the metadata provided to buyers should contain information about the subset utility in the model. We desire a data valuation considering the dataset's upstream cost and downstream utility in NN models from a fairness perspective. Further, the SVBP framework is desired to prevent the value from increasing under dishonest behavior, such as replication and outliers. Finally, we desire the operation of a data market to avoid price discrimination [9] and arbitrage opportunities [10]. Under such constraints, the SVBP is expected to obtain more revenue than FsM with a flat pricing strategy [11].

Our Solution. To address the aforementioned challenges, we take two technical routes, one is a subset valuation method to enable SsM, and the other is to design the SVBP function in order to maximize data vendors' revenue and improve buyers' affordability.

A mature data market is desired to encourage data vendors to provide high-quality data. For NN model tasks, the dataset value is desired to be proportional to the diversity of clusters in a dataset. A high cluster-diversity dataset refers to significant inter-cluster features and rich intra-cluster data points. To this end, we propose a novel data value metric, cluster-based robust volume (*ClusterRV*). We first prove that the data matrix volume (determinant of the matrix's left Gram), which is the basic form of *ClusterRV*, is related to the dataset's utility in the NN model training and verify that volume can be used as a value metric for multi-form subsets, e.g., row subsets and column subsets. In addition, since volume is task-agnostic, the volume-based data valuation avoids price discrimination of assigning different values on the same dataset among multi-tasks. To provide robust data valuation, we use k-Means method in *ClusterRV* to compress the original dataset to a dataset contained clustering centers, which can eliminate the effects of replication and outliers on the dataset's value. Furthermore, we theoretically and experimentally demonstrate the replication-robustness of *ClusterRV* on two metrics, inflation and bounded distortion. Combining *ClusterRV* with the Shapley value, we design *ClusterRVSV* for the equitable data valuation on subsets that satisfies the SsM data market requirements. We carry out extensive experiments to demonstrate the validity of *ClusterRVSV* in various scenarios.

For the revenue maximization, we establish a formal optimization framework of SVBP based on the buyers' WTP curve and demand distribution curve. The arbitrage-free is formalized as a constraint. Sampling on the pricing function, we obtain an optimization problem in the discrete form, where the pricing function can be linearly interpolated from the discrete solution. For the arbitrage-free constraint in sub-additive form, the proof of existence of pricing function is a coNP-hard problem even under a simple revenue model. To address this intractability, we relax the sub-additive constraint and obtain an equivalent optimization problem with approximation guarantee to the original problem. We conduct extensive experiments and show that SVBP provides a higher revenue for data vendors, and a larger affordability ratio, and thus accessibility for more buyers, making the data market more efficient than the existing FsM data market with the flat pricing strategy.

Contributions. Our key technical contribution is to propose a subset-sale mode data market and present algorithmic solutions for its implementation, involving a novel data valuation method and a subset value-based pricing strategy to maximize the revenue.

- To the best of our knowledge, this is the first paper on proposing a formal framework for a neural network data market with a subset-sale mode. We formalize the important properties that SsM should satisfy, such as fair data valuation, avoidance of price discrimination, arbitrage freeness, and so on.
- We investigate the feasibility of using volume as a task-agnostic data valuation metric for arbitrary subsets of datasets by justifying (both theoretically and empirically) its correlation to NN model parameter variations.
- We propose *ClusterRV* to provide robust data valuation under dishonest behavior, such as replication and outliers. Combined with the Shapley value, a novel data valuation method

ClusterRVSV is proposed to satisfy the requirements of data valuation in SsM.
- We formalize an optimization problem to maximize data vendors' revenue with arbitrage freeness as a constraint, where the solution is a feasible SVBP function in SsM.
- Extensive experiments validate that our proposed data valuation is very promising in various scenarios. SVBP provides a higher revenue for data vendors and a larger affordability ratio for buyers than those of the existing FsM data market with the flat pricing strategy.

Roadmap. In Sec. 2, we introduce the concept of volume and derive the relationship between volume and NN model parameter variation (both theoretically and empirically). In Sec. 3, we provide a replication-robust and outlier-robust data valuation method. In Sec.4, we investigate the revenue optimization problem. Sec. 5 conducts the experimental evaluation and result analysis. Sec. 6 briefly discusses the related work. We conclude our paper in Sec. 7.

2 VOLUME-BASED DATA VALUATION

As AI technology advances, researchers become increasingly eager to gather high-quality (high-value) data from data markets for neural networks' training [12–14]. Shapley value is widely used for measuring the data value, and is demonstrated as follows:

$$\psi_{SV}(X_S)) = \sum_{X_T \subset \{X\} \setminus \{X_S\}} \frac{|X_T|!(|X|-|X_T|-1)!(V(X_T \cup X_S)-V(X_T))}{|X|!}, \quad (1)$$

where the value metric V usually adopts the model's validation loss. To obtain the corresponding data value in SVBP for various subset combinations, the validation-loss-based Shapley value (VSV) requires training the model for $|X|!$ times. In practice, buyers even do not provide validation loss to the data market due to commercial confidentiality, making data vendors misvalue their data and set unreasonable prices. In this section, we formalize a measure of data diversity that replaces the validation loss by data's volume. We demonstrate a proposition that a higher volume yields a higher parameter variation during the model training. This proposition holds regardless of the considered row or column subset, laying the groundwork for task-agnostic subset valuation that supports SsM.

2.1 Problem Formulation

Volume. Considering a dataset $D = \{X, Y | X \in R^{N \times M}, Y \in R^{N \times O}\}$, and a full-connected neural network model M, as shown in Figure 2.

The model's training ends at the maximum epochs or an acceptable mean squared error (MSE): $\left\| Y - \widehat{Y} \right\|^2 \leq \epsilon$, and at this time, the model parameter for each layer is written as follows:

$$y_{in}^1 := Xw^1 \rightarrow w^1 := X^+ y_{in}^1. \quad (2)$$

Considering linear activation functions $\{\sigma^i\}$ among layers, the w_2 can be calculated as follows,

$$y_{in}^2 := \sigma(y_{in}^1)w^2 = y_{in}^1 w^2 = Xw^1 w^2 \rightarrow w^2 := \{w^1\}^+ X^+ y_{in}^2. \quad (3)$$

Up to layer l_L,

$$w^L := \prod_{i=1}^{L-1} \{w^i\}^+ X^+ y_{in}^L. \quad (4)$$

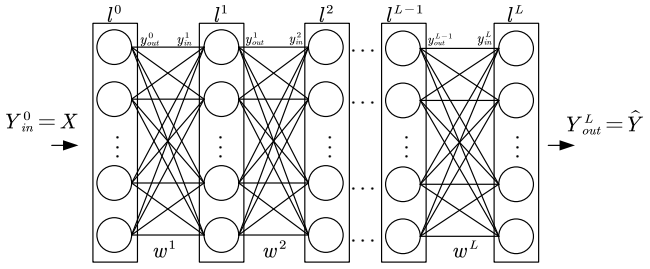

Figure 2: The schema of a full-connected neural network with L layers, where y_{in}^i and y_{out}^i are the input and output of the layer l_i after training. The parameter of each layer i in the trained model is denoted as w^i. $\widehat{Y}$ is the trained model's output corresponding to input X.

The above derivation shows a theoretical connection between X^+ and parameter W, allowing a volume-based data valuation. In the following section, we give a theoretical analysis of a volume-based data valuation to assign a higher value for data that leads to a higher parameter variation. Denoting the left Gram matrix of X is $G := X^T X \in R^{M \times M}$ and $|X| := det(X)$, the definition of **Volume** is as follows:

Definition 1. ***(Volume)*** *The volume of a matrix X can be straightforwardly generalized to $\sqrt{\left|(X^T X)\right|}$ or $\sqrt{\left|(XX^T)\right|}$, depending on the shape of X. Assume a full-rank matrix $X \in R^{N \times M}$ and $M < N$:*

$$Vol(X) := \sqrt{\left|(X^T X)\right|} = \sqrt{|G|}. \quad (5)$$

We adopt $Vol(X) := \sqrt{\left|(X^T X)\right|}$ for the number of features (columns) is often much less than the number of data points (rows). In addition, the feature space of data is often predetermined and fixed while the data points grow indefinitely by incrementally updating. Except for the connections to parameters, volume is proportional to data size. Adding new data points to a dataset will increase its volume.

Lemma 1. *For a invertible matrix A and any column vector u and v, then there exits:*

$$\det\left(A + uv^T\right) = \det(A)\left(1 + v^T A^{-1} u\right). \quad (6)$$

Proposition 1. *Considering a matrix $\bar{X} = \left[X^T, x^T\right]^T$, where $X \in R^{N \times M}$ and $x \in R^{1 \times M}$, $Vol(\bar{X}) \geq Vol(X)$ is hold.*

Proof 1. ***(Proposition 1)*** *According to the property of matrix determinant lemma (Lemma 1), the square volume of $\bar{X}$ can be written as:*

$$\begin{aligned} Vol(\bar{X})^2 &= \left|\bar{X}^T \bar{X}\right| = \left|\left[X^T, x^T\right]\begin{bmatrix} X \\ x \end{bmatrix}\right| \\ &= \left|X^T X + x^T x\right| = \sigma\left|X^T X\right| = \sigma Vol(X)^2, \end{aligned} \quad (7)$$

where $\sigma = 1 + x(X^T X)^{-1} x^T \geq 1$. Thus, $Vol(\bar{X}) \geq Vol(X)$ is hold and adding new data points to a dataset will increase its volume is proved.

In this sense, volume accounts for the vendor valuing preference, which is datasize and is related to the buyer valuing preference, that is, data utility in model training. At the same time, volume can be efficiently applied to various data subsets. As a result, volume is suitable as a data value metric in the SsM data market.

Note. Full rank is a strict constraint in Equation 5, otherwise, $Vol(X) \equiv 0$. To satisfy this constraint, dimensionality reduction techniques can be used to remove linearly dependent columns, such as Gram-Schmidt process [15], principal component analysis [16], and independent components analysis [17].

2.2 A Larger Volume Leads to A Larger Parameter Variation

To validate the previously stated observation that volume is related to the model parameters, we first introduce the parameter variation and its relationship to bias. Second, we give propositions indicating that a larger volume leads to a larger bias. Finally, we propose a corollary that a larger volume results in a larger parameter variation.

Parameter Variation and Bias. Considering a neural network with parameter W and a training dataset $D = \{X, Y\}$, where $X \in R^{N\times M}, Y \in R^{N\times O}$, the training goal is to minimize the MSE loss: $J(X; W) := \left\|Y - \hat{Y}(X; W)\right\|^2$. $\hat{Y}(X; W)$ corresponds the output of the model. Assuming using mini-batch gradient descent to update parameters, and setting $batch\_size = 1$, the parameter variation caused by data point $x_i \in X$ is:

$$\triangle W_i := -\alpha \bigtriangledown_W J(x_i; W_{i-1}) = -\alpha(Y - \hat{Y}(x_i; W_{i-1})) \bigtriangledown \hat{Y}(x_i; W_{i-1}). \tag{8}$$

In Equation 8, two factors influence the parameter variation: 1) the output value of the model, $\hat{Y}$; 2) the gradient of the model's output, $\hat{Y}(X; W)$. As a result, a higher parameter variation at the input x_i corresponds to a higher variation in the model's output or gradient. In scenarios that do not involve the market [18, 19], the data owner is also the model owner. Thus, Shapley value (Equation 1) of a data point x_i can be performed on the model's loss directly. In the data market, where loss functions vary among buyers, the parameter variation provides an alternative substitution of the model's loss to reflect data utility in model training. Buyers expect new data to bring significant changes to the original model, and such expectations are consistent. However, as the commercial secrets of buyers, neither loss nor parameter variation can be obtained by data vendor and market. To demonstrate the relationship between the data volume and the model parameter variation, we convert the fore-mentioned parameter variation to matrix form and introduce matrix bias.

Definition 2. ***(Bias of Submatrix)*** *For any $X \in R^{N\times M}$, and X_i are subsets in X. X^+ and X_i^+ is the corresponding pseudo-inverse of X and X_i, respectively. The bias of X_i is defined as the L2-norm of the difference between X^+ and X_i^+:*

$$bias := \left\|X^+ - X_i^+\right\|. \tag{9}$$

Considering the row submatrices $X_1, X_2 \in R^{\frac{N}{2}\times M}$ with full columns and $X = \left[X_1^T, X_2^T\right]^T \in R^{N\times M}$. As shown in figure 2, the layer l's input of a NN trained by X_1 or X_2 are respectively denoted as y_1^l and y_2^l. According to Equation 4, the layer l's parameter w^l trained by X_1, X_2 and X respectively are:

$$\begin{aligned} &w_1^l = \prod_{i=1}^{l-1} \left\{w_1^i\right\}^+ \{X_1\}^+ y_1^l,\ w_2^l = \prod_{i=1}^{l-1} \left\{w_2^i\right\}^+ \{X_2\}^+ y_2^l, \\ &w^l = \prod_{i=1}^{l-1} \left\{w^i\right\}^+ X^+ y^l = \prod_{i=1}^{L-1} \left\{w^i\right\}^+ \left[\left\{X^+\right\}_1 \quad \left\{X^+\right\}_2\right] \begin{bmatrix} y_1^l \\ y_2^l \end{bmatrix}. \end{aligned} \tag{10}$$

To measure the independent effect of the addition of each $X_j, j \neq i$ on the the model with parameter $\left\{w_{X_i}^l, l = 1, ..., L\right\}$, which have trained on X_i, we adopt control variate method and set $y_j^l = 0, j \neq i$. Under a front-to-back layer-by-layer calculation, the parameters of fore-layers are considered as fixed constant vector. The approximate average parameter variation on L layers,$variation_j$, induced by X_j is defined as follows:

$$\begin{aligned} &variation_j := \|W(X) - W(X_i)\| = \frac{1}{L}\sum_{l=1}^{L} \left\|w^l - w_i^l\right\| \\ &\propto \frac{1}{L}\sum_{l=1}^{L} \left\|\left\{X^+\right\}_i y_i^l - \{X_i\}^+ y_i^l\right\| \propto \left\|\left\{X^+\right\}_i - \{X_i\}^+\right\| = bias_i. \end{aligned} \tag{11}$$

Similarly, for column submatrices $X_1, X_2 \in R^{N\times\frac{M}{2}}$ with full rows and $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \in R^{N\times M}$, the layer l's parameter w^l trained by X_1, X_2 and X respectively are:

$$\begin{aligned} &w_1^l = \prod_{i=1}^{L-1} \left\{w_1^i\right\}^+ \{X_1\}^+ y^l,\ w_2^l = \prod_{i=1}^{L-1} \left\{w_2^i\right\}^+ \{X_2\}^+ y^l, \\ &w^l = \prod_{i=1}^{L-1} \left\{w^i\right\}^+ X^+ y^l = \prod_{i=1}^{L-1} \left\{w^i\right\}^+ \begin{bmatrix} \left\{X^+\right\}_1 \\ \left\{X^+\right\}_2 \end{bmatrix} y^l. \end{aligned} \tag{12}$$

Using the same assumptions as above, the approximate average parameter variation on L layers, $variation_j$, induced by a column subset X_j is defined as follows:

$$\begin{aligned} &variation_j := \|W(X) - W(X - X_i)\| = \frac{1}{L}\sum_{l=1}^{L} \left\|w^l - w_i^l\right\| \\ &\propto \frac{1}{L}\sum_{l=1}^{L} \left\|\left\{X^+\right\}_i y^l - \{X_i\}^+ y^l\right\| \propto \left\|\left\{X^+\right\}_i - \{X_i\}^+\right\| = bias_i. \end{aligned} \tag{13}$$

A Smaller Bias Leads to A Larger Parameter Variation. The above relation can be extended to any subset X_s; that is, the variation in parameters introduced by subset X_s is proportional to the bias between the full set X and the complement of a subset X_s in X: $\{X \setminus X_s\}$. This conclusion is intuitive that as an end-to-end model, the parameter variation in the neural network is only related to the input variation. The matrix concept *bias* estimates the geometric distance between $\{X \setminus X_s\}^+$ and X^+. When a model has been trained on a dataset $\{X \setminus X_s\}$ that is much similar to X, the model output is closer to the target's output. When a new training dataset X_s is appended, it can generate less gradient and less parameter variation.

A Larger Volume Leads to A Smaller Bias. To infer that a larger volume leads to a larger parameter variation, we refer to the following statement about the relationship between volume and bias [20] and demonstrate cases with interpretability where a large volume leads to a smaller bias in row subsets and column subsets, respectively.

Proposition 2. ***(Volume w.r.t. Bias for** rank = 1**)** For a matrix $X \in R^{N\times 1}$, and its row submatrices $X1, X2 \in R^{\frac{N}{2}\times 1}$ or a matrix $X \in R^{N\times 2}$ and its column submatrices $X1, X2 \in R^{N\times 1}$, if $Vol(X_1) \geq Vol(X_2)$, then $bias_1 \leq bias_2$.*

Proof 2. ***(Proposition 2)** For any two sub-matrix : $X_1, X_2 \in R^{N\times 1}$, regardless of row or column sub-matrix of the original matrix. Note for any matrix X_S with rank = 1, $\left(\mathbf{X_S}^\top \mathbf{X_S}\right)^{-1} = \|\mathbf{X_S}\|^2 = Vol(X_S)^2$. Considering the pseudo-inverse is denoted as: $\mathbf{X_S}^+ := \left(\mathbf{X_S}^\top \mathbf{X_S}\right)^{-1} \mathbf{X_S}^\top$, the bias variation is:*

$$\begin{aligned}
&bias_2^2 - \ bias_1^2 = \left\|X^+ - X_2^+\right\|^2 - \left\|X^+ - X_1^+\right\|^2 \\
&= \left\|\mathbf{X}^+\right\|^2 - 2\left\langle \mathbf{X}^+, \mathbf{X}_2^+\right\rangle + \left\|\mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}^+\right\|^2 + 2\left\langle \mathbf{X}^+, \mathbf{X}_1^+\right\rangle - \left\|\mathbf{X}_1^+\right\|^2 \\
&= \left\|\mathbf{X}^+\right\|^2 - 2\left\|\mathbf{X}_2^+\right\|^2 + \left\|\mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}^+\right\|^2 + 2\left\|\mathbf{X}_1^+\right\|^2 - \left\|\mathbf{X}_1^+\right\|^2 \\
&= \left\|\mathbf{X}_1^+\right\|^2 - \left\|\mathbf{X}_2^+\right\|^2 = \frac{1}{\|\mathbf{X_1}\|^4}\left\|\mathbf{X_1}^\top\right\|^2 - \frac{1}{\|\mathbf{X_2}\|^4}\left\|\mathbf{X_2}^\top\right\|^2 \\
&= \frac{1}{\|\mathbf{X_1}\|^2} - \frac{1}{\|\mathbf{X_2}\|^2} = \frac{1}{\mathrm{Vol}\left(\mathbf{X_1}\right)^2} - \frac{1}{\mathrm{Vol}(\mathbf{X_2})^2}.
\end{aligned} \tag{14}$$

Since the $Vol(X_1)$ and $Vol(X_2)$ are constants, a larger volume yields a smaller bias is proved.

Proposition 3. ***(Volume w.r.t. Bias for** rank > 1**)** For a matrix $X \in R^{N\times M}$, and its row submatrices $X1, X2 \in R^{\frac{N}{2}\times M}$ or column submatrices $X1, X2 \in R^{N\times \frac{M}{2}}$, the following relation can be constructed by Sylvester's formula :*

$$\begin{aligned}
&bias_2^2 - bias_1^2 = \left\|X^+ - X_2^+\right\|^2 - \left\|X^+ - X_1^+\right\|^2 \\
&= \frac{\left\|Q_2 X_2^T\right\|^2}{Vol(X_2)^4} - \frac{\left\|Q_1 X_1^T\right\|^2}{Vol(X_1)^4} + 2\left\langle \frac{QX^T}{Vol(X)^2}, \frac{Q_1 X_1^T}{Vol(X_1)^2} - \frac{Q_2 X_2^T}{Vol(X_2)^2}\right\rangle,
\end{aligned} \tag{15}$$

where,

$$Q = \textstyle\sum_{l=1}^{k}\left((\lambda_l\sigma_l)^{-1}\prod_{j=1,j\neq l}^{k}(G-\lambda_i I)\right), \tag{16}$$

$$\begin{aligned}
&\sigma_l = \textstyle\sum_{g=1}^{k}(-1)^{g+1}\cdot \\
&\left(\lambda_l^{k-g}\textstyle\sum_{\mathcal{H}\subseteq\{1,\ldots,k\}\setminus\{l\},|\mathcal{H}|=g-1}\left(\prod_{h\in\{1,\ldots,k\}\setminus\mathcal{H}}\lambda_h^{-1}\right)\right).
\end{aligned} \tag{17}$$

$\{\lambda_l, l = 1, ..., k\}$ denotes the left Gram matrix G's k unique eigenvalues.

Proof 3. ***(Proposition 3)** The proposition 2 is expanded by substituting $\mathbf{X}^+ = \mathbf{G}^{-1}\mathbf{X}^\top$, where $\mathbf{G} := \mathbf{X}^\top\mathbf{X}$ is the left Gram matrix.*

$$\begin{aligned}
&\mathrm{bias}_2^2 - \mathrm{bias}_1^2 := \left\|\mathbf{X}^+ - \mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}^+ - \mathbf{X}_1^+\right\|^2 \\
&= \left\|\mathbf{X}^+\right\|^2 - 2\left\langle \mathbf{X}^+, \mathbf{X}_2^+\right\rangle + \left\|\mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}^+\right\|^2 + 2\left\langle \mathbf{X}^+, \mathbf{X}_1^+\right\rangle - \left\|\mathbf{X}_1^+\right\|^2 \\
&= \left\|\mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}_1^+\right\|^2 + 2\left\langle \mathbf{X}^+, \mathbf{X}_1^+ - \mathbf{X}_2^+\right\rangle \\
&= \left\|\mathbf{G}_2^{-1}\mathbf{X}_2^\top\right\|^2 - \left\|\mathbf{G}_1^{-1}\mathbf{X}_1^\top\right\|^2 + 2\left\langle \mathbf{G}^{-1}\mathbf{X}^\top, \mathbf{G}_1^{-1}\mathbf{X}_1^\top - \mathbf{G}_2^{-1}\mathbf{X}_2^\top\right\rangle,
\end{aligned} \tag{18}$$

where $\mathbf{G}^{-1}$ can be expanded by Sylvester's matrix theorem.

Theory 1. ***(Sylvester's Matrix Theorem)** Given a diagonalizable square matrix X, any analytic function $f(\cdot)$ can be expanded,*

$$f(\mathbf{X}) = \sum_{l=1}^{k} f\left(\lambda_l\right)\mathbf{X}_l, \tag{19}$$

where λ_l is the i-th distinct eigenvalue of X and X_l is the Frobenius covariant as follows:

$$\mathbf{X}_l := \prod_{j=1,j\neq l}^{k}\frac{1}{\lambda_l - \lambda_j}\left(\mathbf{X} - \lambda_j\mathbf{I}\right). \tag{20}$$

Corollary 1. *Supposing $f(\cdot)$ is a inverse function, then,*

$$f(\mathbf{X}) = \mathbf{X}^{-1} = \sum_{l=1}^{k}\frac{1}{\lambda_l}\mathbf{X}_l. \tag{21}$$

For $\mathbf{G}$ is diagonalizable matrix, $\mathbf{G}^{-1}$ can be written as follows by above corollary.

$$\mathbf{G}^{-1} = \sum_{l=1}^{k}\frac{1}{\lambda_l}\mathbf{G}_l = \sum_{l=1}^{k}\frac{1}{\lambda_l}\left(\prod_{j=1,j\neq l}^{k}\frac{1}{\lambda_l - \lambda_j}\prod_{j=1,j\neq l}^{k}\left(\mathbf{G} - \lambda_j\mathbf{I}\right)\right). \tag{22}$$

Denoting $p_l = \prod_{j=1,j\neq l}^{k}\frac{1}{\lambda_l - \lambda_j}$ and expanding the products, we have:

$$\frac{1}{p_l} = \underbrace{\Lambda\sum_{g=1}^{k}(-1)^{g+1}\lambda_l^{k-g}\left[\sum_{\mathcal{H}\subseteq\{1,\ldots,k\}\setminus\{l\},|\mathcal{H}|=g-1}\left(\prod_{h\in\{1,\ldots,k\}\setminus\mathcal{H}}\frac{1}{\lambda_h}\right)\right]}_{\sigma_l}, \tag{23}$$

where $\Lambda = \prod_{i=l}^{k}\lambda_l = |\mathbf{G}|$. The Equation 22 can be rewritten as:

$$\begin{aligned}
\mathbf{G}^{-1} &= \frac{1}{|\mathbf{G}|}\underbrace{\sum_{l=1}^{k}\frac{1}{\lambda_l}\frac{1}{\sigma_l}\prod_{j=1,j\neq l}^{k}\left(\mathbf{G} - \lambda_j\mathbf{I}\right)}_{\mathrm{Q}} \\
&= \frac{1}{Vol(X)^2}\underbrace{\sum_{l=1}^{k}\frac{1}{\lambda_l}\frac{1}{\sigma_l}\prod_{j=1,j\neq l}^{k}\left(\mathbf{G} - \lambda_j\mathbf{I}\right)}_{\mathrm{Q}}.
\end{aligned} \tag{24}$$

Substituting $\mathbf{G}^{-1}(\mathbf{G_1}^{-1}, \mathbf{G_2}^{-1})$ in Equation 18 by above result, the Equation 18 can be rewritten as:

$$\begin{aligned}
\mathrm{bias}_2^2 - \mathrm{bias}_1^2 = &\frac{\left\|Q_2 X_2^T\right\|^2}{Vol(X_2)^4} - \frac{\left\|Q_1 X_1^T\right\|^2}{Vol(X_1)^4} \\
&+ 2\left\langle \frac{QX^T}{Vol(X)^2}, \frac{Q_1 X_1^T}{Vol(X_1)^2} - \frac{Q_2 X_2^T}{Vol(X_2)^2}\right\rangle.
\end{aligned} \tag{25}$$

When $rank = 1$, due to full rank assuming, the column feature dimension is 1. Thus, the **Proposition 2** is easy to be proof for that the left Gram matrix is a scalar. For $rank > 1$, the **Proposition 3** is hold under two cases: (1) **Case 1**: $\left\|Q_1 X_1^T\right\|^2 \approx \left\|Q_2 X_2^T\right\|^2$ and $Vol(X) \gg max(Vol(X_1), Vol(X_2))$; (2) **Case 2**: $Vol(X_1) \gg Vol(X_2)$. For row subset $X_1, X_2 \in X$, one condition $Vol(X) \gg max(Vol(X_1), Vol(X_2))$ in **Case 1** is commonly hold and is proved in **Proposition 4**. Another condition $\left\|Q_1 X_1^T\right\|^2 \approx \left\|Q_2 X_2^T\right\|^2$ is hold when X_1 and X_2 are similar. Due to the data points follow independent and identical distribution (i.i.d.) in a dataset, **Case 1** widely exists in row subsets. But for column subsets $X_1, X_2 \in X$ containing distinct features, both $Vol(X) \gg max(Vol(X_1), Vol(X_2))$ and $\left\|Q_1 X_1^T\right\|^2 \approx \left\|Q_2 X_2^T\right\|^2$ conditions are not guaranteed. From

a perspective that $Vol()$ measures the diversity in the features [21], a high volume difference exists between $X1$ and $X2$ when they follow significant different distributions. In such $X1$ and $X2$, $Vol(X_1) \gg Vol(X_2)$ or $Vol(X_2) \gg Vol(X_1)$ is hold. Thus, for column subsets, **Case 2** exists more commonly.

Proposition 4. *For any matrix $X = \left[X_1^T, X_2^T\right]^T \in R^{N\times M}$ and $X_1^T, X_2^T \in R^{\frac{N}{2}\times M}$ are submatrices in row, there exists $Vol(X) \gg max(Vol(X_1), Vol(X_2))$ when $N \to \infty$.*

Proof 4. ***(Proposition 4)*** *Substituting the vector x in **Proposition 1** by a matrix $X' = \left[x_1^T, ..., x_n^T\right]^T$ and adding row vectors of X' to matrix X row by row, the following equation exists.*

$$\begin{aligned} Vol\left(\left[X^T, X'^T\right]^T\right)^2 &= Vol\left(\left[X^T, x_1^T, ..., x_n^T\right]^T\right)^2 \\ &= \sigma_n Vol\left(\left[X^T, x_1^T, ..., x_{n-1}^T\right]^T\right)^2 = ... = \prod_{i=1}^{n} \sigma_i Vol(X)^2, \end{aligned} \tag{26}$$

where $\sigma_i = 1 + x_i(X^TX)^{-1}x_i^T \geq 1$ and $\prod_{i=1}^n \sigma_i \to \infty$ when $n \to \infty$. Then,

$$Vol\left(\left[X^T, X'^T\right]^T\right) >> Vol(X), \tag{27}$$

***Proposition 4** can be proved by taking X_1 and X_2 as a new addition submatrix, respectively.*

We empirically verify the aforementioned analysis. Figure 3 and Figure 4 show the percentage of whether a larger volume yields smaller bias in row subsets and column subsets, respectively. All experiments are performed for 100 trials. In each trial, X_1 and X_2 are two equal-sized subsets splited from the full-set $X \in R^{N\times M}$. The row size N ranges in $100 \sim 10000$ and the column size M ranges in $[1, 2, 5, 10]$. In the row subsets experiment (Figure 3), $X_1 \in R^{\frac{N}{2}\times M}$, $X_2 \in R^{\frac{N}{2}\times M}$ follows i.i.d. of $\mathcal{N}(0,1)^M$ (left graph) and $\mathcal{U}(0,1)^M$ (right graph), respectively. When $M = 1$, the percentage of a larger volume yields smaller bias is strictly 100%, which consistent with **Proposition 1**. When $M > 1$, for X_1, X_2 following i.i.d., experimental result shows that the probability of a larger volume yields smaller bias is greater than 80%. In the colume subset experiment (Figure 4), the subset size is: $X_1 \in R^{N\times\frac{M}{2}}$ and $X_2 \in R^{N\times\frac{M}{2}}$. When X_1 and X_2 follows the same distribution $\mathcal{N}(0,1)^M$ (left graph), the probability of a larger volume yields smaller bias under $M > 1$ is greater than 60% (**Case 1**). When X_1 and X_2 follows distinguished feature distribution of $\mathcal{N}(0,1)^M$ and $\mathcal{U}(0,1)^M$ respectively (right graph), the probability of a larger volume yields smaller bias under $M > 1$ is equal to 100%, for in this case, $Vol(X_1) \gg Vol(X_2)$ (**Case 2**). The experiment results are consistent with what we expect in **Proposition 2, 3**. Combining the above propositions with Equations (11,13), we give the inference about the relation between the volume and the NN's parameter variation as follows.

Corollary 2. *For a matrix $X \in R^{N\times M}$, and its row submatrix $X1, X2 \in R^{\frac{N}{2}\times M}$ or column submatrix $X1, X2 \in R^{N\times\frac{M}{2}}$, if $Vol(X_1) \geq Vol(X_2)$, $variation_1 \geq variation_2$.*

The above relationship between the volume and the parameter variation in NN model training is further verified in Sec. 5. As a first step toward an SsM data market, in this section, we focus on developing a task-agnostic data value metric, volume, that applies to subsets in various forms. We conclude this section with a volume-based Shapley value (VSV) for equitable data valuation on subsets in SVBP. The VSV is defined as follows for any submatrices X_S in X.

$$\psi_{VSV}(X_S) = \sum_{X_T \subset \{X\}\setminus\{X_S\}} \left(\frac{|X_T|!(|X|-|X_T|-1)!(Vol(X_T \cup X_S) - Vol(X_T))}{|X|!} \right). \tag{28}$$

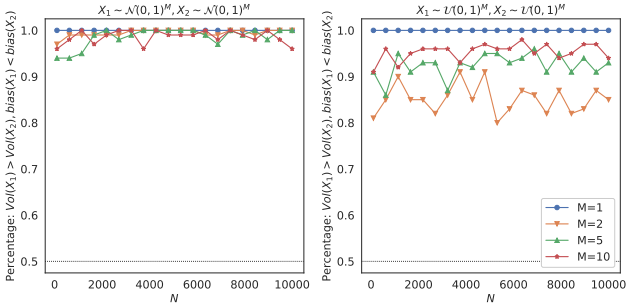

Figure 3: Volume vs. bias for two equal-sized X_1 and X_2, which are identically sampled from the rows of $X \in R^{N\times M} \sim \mathcal{N}(0,1)^M$ (left) and $X \in R^{N\times M} \sim \mathcal{U}(0,1)^M$ (right), respectively. The ordinate axis shows the percentage of times whether a larger volume leads to a smaller bias under 100 trials.

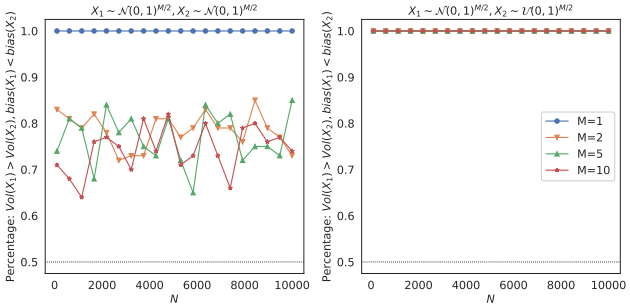

Figure 4: Volume vs. bias for two equal-sized X_1 and X_2, which are sampled from the columns of $X \in R^{N\times M}$, and with identical (left) and different distributions (right), respectively. The ordinate axis shows the percentage of times whether a larger volume leads to a smaller bias under 100 trials.

3 ROBUSTNESS DATA VALUATION FOR REPLICATION AND OUTLIERS

Definition 3. ***(Replication Robustness)*** *Replication robustness can be measured by value inflation ratio: $\gamma_\nu := \frac{sup_{c\geq 1}\, \nu(replicate(X,c))}{\nu(X)}$, where $\nu()$ is a data value metric function that maps a dataset X to a real value, $replicate(X, c)$ is a replication function that copies X for c times. The optimal replication robustness exists when $\gamma_\nu = 1$.*

The aforementioned VSV (Equation 28) enables task-agnostic data valuation for row subsets and column subsets. However, it is not with replication-robustness and outlier-robustness.

Proposition 5. ***(Volume is not robust to replication)*** *For a N-by-M dataset X, a replicated $(N + |X_S|\, d)$-by-M dataset $X_{rep} := \left[X^T, X_S^T, ..., X_S^T\right]^T$ is generated when replicating a row subset $X_s \in X$ for $d > 0$ times. Volume is not robust to replication for $Vol\left(X_{rep}\right) > Vol\,(X)$ and $\lim_{d\to\infty} Vol\left(X_{rep}\right) = \infty$.*

Proof 5. ***(Proposition 5)*** *Considering a replication-involving dataset $X_{rep} = \left[X^T, \underbrace{X_S^T, ..., X_S^T}_{d}\right]^T$, where X_S is row vector and is replicated for d times. According to Equation 26, the square volume of X_{rep} is written as:*

$$Vol(X_{rep})^2 = (1 + X_S(X^TX)^{-1}X_S^T)^d \left|X^TX\right|. \tag{29}$$

For $(1 + X_S(X^TX)^{-1}X_S^T) > 1$, the exponential increasing of volume under replication is proved and $lim_{d\to\infty} Vol\left(X_{rep}\right) = \infty$ is hold.

Proposition 6. ***(Volume is not robust to outliers)*** *For a $(N + d)$-by-M dataset $X_{outlier} := \left[X^T, outlier_1^T, ..., outlier_d^T\right]^T$, where $outlier_i \in R^{1\times M}$, volume is not robust to outliers for $Vol\,(X_{outlier}) > Vol\,(X)$ and $\lim_{d\to\infty} Vol\,(X_{outlier}) = \infty$.*

Proof 6. ***(Proposition 6)*** *Considering a outlier-injection dataset $X_{outlier} = \left[X^T, \underbrace{outlier_1^T, ..., outlier_d^T}_{d}\right]^T$, where $outlier_i$ is row vector and is replicated for d times. According to Equation 26, the square volume of X_{rep} is written as:*

$$Vol(X_{outlier})^2 = \left(\prod_{i=1}^{d}(1 + outlier_i(X^TX)^{-1}outlier_i^T)\right)\left|X^TX\right|. \tag{30}$$

For $(1+outlier_i(X^TX)^{-1}outlier_i^T) > 1$, the exponential increasing of volume under replication is proved and $lim_{d\to\infty} Vol\,(X_{outlier}) = \infty$ is hold.

In a geometric perspective, the volume shows how many times an M-dimensional Euclidean volume of an image of a unit N-dimensional ball under a linear operator X^T is greater than the Euclidean volume of a unit M-dimensional ball for a N-by-M matrix X with $N \geq M$. Therefore, adding the replicated data point or outliers to X will monotonically increase the volume. For replication-robustness, a most natural way is to remove them. For a trade-off between removing replication and reserving data diversity, [20] propose a discretization-based volume (**Method 1**).

Method 1. ***(Discretization-based Robust Volume)*** *Given a discretization coefficient ω, the input domain for X is discretized into a set of cubes with sides of length ω. Let ψ denote the set of indices of these d-cubes, and ϕ_i denote the number of data points in the i-th cube. The discretization-based robust volume is defined as follows.*

$$\omega RV(X) = Vol(\tilde{X}) \times \prod_{i\in\Psi} \rho_i, \tag{31}$$

where $\rho_i := \sum_{p=0}^{\phi_i} \alpha^p$, $\tilde{X} := \{mean(x_i)|\phi_i \neq 0, i \in \Psi\}$ is a compressed version of X where $mean(x_i)$ is a statistic of the data points in the i-th cube, and $\alpha \in (0, 1)$ controls the trade-off between replication robustness and diversity reservation.

An intractable problem in **Method 1** is how to determine an appropriate discretization coefficient ω, especially for a dataset with feature columns distributed in various scales. Thus, the adaptability of a fixed ω, in this case, is limited. Another question worth considering is outlier-sensitivity. Outliers are sporadic, random data caused by sensor failures, manual entry errors, or unusual events; thus, zero value should be allocated. The discretization-based mechanism enables an outlier in its specific cube and thus be retained. Considering a outlier x_o, its relative value cannot be excluded by *mean* on a specific cube in **Method 1**, even promoted from $Vol(x_0)/Vol(X)$ to $Vol(x_0)/Vol(\tilde{X})$ after data compression. Another question should be considered in a NN model-faced data market is that, intuitively, a dataset with higher cluster diversity corresponds to a higher-quality and thus higher value than a same-size dataset but with lower cluster diversity. In a MNIST handwritten digit classification problem, for example, subset 1 contains 100 randomly sampled handwritten pictures ranging from clusters labeled "0" to "9", while subset 2 contains 200 handwritten pictures sampled from clusters labeled "0" to "5". Compared to subset 1, subset 2 is with higher data diversity for containing more unique data and lower cluster diversify for sampling from 5 clusters. Figure 5 shows that a model trained on subset 1 is with better classification performance than subset 2, meaning that a higher cluster diversity is more valuable to indicate better learning performance. To address

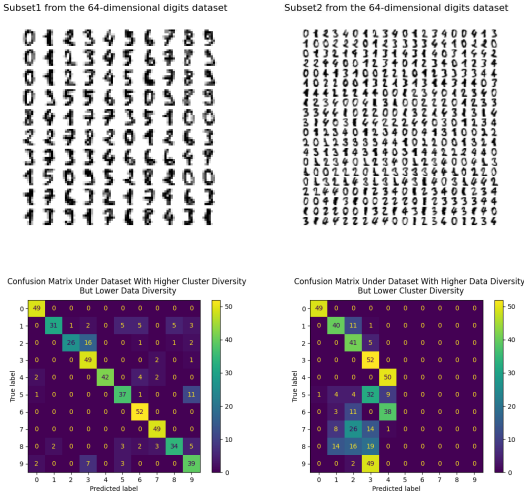

Figure 5: Confusion matrix under different training datasets.

the aforementioned issues, we propose the cluster-based robust volume (**Method 2**).

Method 2. ***(Cluster-based Robust Volume)*** *Given the number of clusters K, the dataset is partitioned into K distinct non-overlapping clusters by k-Means (Algorithm 1). Let ϕ_i denotes the number of data points in the i-th cluster and x_{center_i} denotes the cluster center of the*

Algorithm 1: Implementation details of k-Means

Data: Dataset: X, Max number of clusters: K, Termination condition: ϵ;
Result: Cluster centers X_c;
Initialise $Clusters = \{Cluster_i\}, Cluster_i = \{\emptyset\}, i = 1, ..., K$;
Randomly initialise cluster centers $X_c^0 = \{c_1, c_2, ..., c_K\}$;
Set iteration index $k = 0$;
while $k == 0$ *or* $\left\|X_c^k - X_c^{k-1}\right\| > \epsilon$ **do**
  for *each data points* x_i *in* X **do**
    Calculate the Euclidean distance:
    $d_i = \arg\min_i \|x_i - c_i\|$;
    Assign x_i to $Cluster_i$;
  end
  Update clusters:
  $X_c^{k+1} \longleftarrow \left\{\frac{1}{|Cluster_i|}\sum_{x \in Cluster_i} x\right\}, i = 1, ..., K$;
  $k \longleftarrow k + 1$;
end
if singleton cluster exist **then** assign it to the closest cluster;
Return: X_c^k

i-th cluster. The cluster-based robust volume is defined as follows.

$$ClusterRV(X) = Vol(X_C) \times \prod_{i \in K} \rho_i, \tag{32}$$

where $\rho_i := \sum_{p=0}^{\phi_i} \alpha^p$, $X_C = [x_{center_1}^T, ..., x_{center_K}^T]^T$ *is a matrix composed of clusters' centers.* $\alpha > 0$ *is trade-off coefficient to balance the degree of data compression and data diversity.*

Different with **Method 1**, which discretizes each dimension by a fixed ω, $ClusterRV$ assigns instances to their closest center and recomputes the centers until the summation of intra-cluster variance reaching at a local minimum. Thus, the cluster-based volume is more scalability when features are distributed in various scales. Only the cluster centers are retained in the compressed matrix X_C, making $ClusterRV(X)$ formalize clusters diversity as data value metric. The diversity of data points is reserved via the weighted term $\prod_{i \in K} \rho_i$ related to the number of data points in each cluster.

Proposition 7. *For a dataset* $X \in R^{N \times M}$, *ClusterRV is with outlier robustness, for* $N \to \infty$,

$$ClusterRV(\{X, x_{outlier}\}) = ClusterRV(X) \tag{33}$$

The proof of **Proposition 7** relies on the mechanism that a singleton cluster will be merged to another nearest non-singleton cluster in k-Means, instead of being separated as a cube in **Method 1**. When the dataset size $N \to \infty$, the effect of sporadic outliers on cluster center is tending to 0. The limitation of k-Means is the cluster solution dependent on the choice of cluster number K and initial centers [22]. Bad initialization can lead to significantly varying clustering results in each restart, making $RV_{cluster}(X)$ varying, particularly for problems with a large search space. One potential solution in a data market is to incent the data vendors to provide metadata containing the clustering information, such as how many classes the data can be clustered into. When metadata is missing,

Algorithm 2: Implementation details of fast global MinMax k-Means

Data: Dataset: X, Parameter: $\theta \in (0, 1)$, the number of clusters: K
Result: Cluster centers X_c
Initialise the centroid c_0 of the data set X:
$c_0 = \frac{1}{|X|}\sum_{i=1}^{|X|} x_i, x_i \in X$, $X_c^0 = \{c_0\}$;
Set iteration index $k = 1$;
while $k \leq K$ **do**
  while *1* **do**
    Set $X_c = \{c_0, c_1, ..., c_{k-1}\}$;
    for *each data points* $x_i \in X$ **do**
      Calculate squared distance between x_i and the closest center among the $k-1$ cluster centers:
      $d_{k-1}^i = min\left\{\left\|x_i - c_j\right\|^2\right\}, j = 0, 1, ..., k-1$
    end
    for *each data points* $x_i \in X$ **do**
      Calculate the quantity:
      $b_i = \sum_{j=1}^{|X|} max(d_{k-1}^j - \left\|x_i - x_j\right\|^2, 0)$
    end
    Set data point x_m with maximum $b_m \in \{b_i\}$ as a starting point for the k-th cluster center,
    $X_c = \{c_0, c_1, ..., c_{k-1}, x_m\}, m = \arg\max_m \{b_m\}$;
    Apply the k-Means algorithm with X_c as initial clusters and get new cluster centers X_c';
    if singleton cluster exist **then** $b_m \longleftarrow 0$ **else** break;
  end
  if all b_i is 0 **then** Return: X_c^{k-1} **else** $X_c^k \longleftarrow X_c'$;
  $k \longleftarrow k + 1$;
end
Return: X_c^K.

the optimal K can be set by linear search with a optimal function of internal clustering validation measures [23] or adaptive k-Means regardless of K value, such as ISODATA [24] and Minmax-distance (MMD) [25]. To produce optimal clustering solution, K-means++ [26, 27] and MinMax k-Means [22], is competitive clustering method insensitive to clusters initialization. In the experiments of Sec. 5, we employ modified global MinMax k-Means [28] as cluster method and the related pseudo code is shown in Algorithm 2.

Next, we use inflation ratio and bounded distortion suggested in [20] to assess the replication robustness of the proposed method.

$$inflation = \frac{ClusterRV(replicate(X, c))}{ClusterRV(X)}, \tag{34}$$

$$bounded\_distortion = \frac{ClusterRV(replicate(X_1, c))V(X_2)}{V(X_1)ClusterRV(replicate(X_2, c))}, \tag{35}$$

where X_1 and X_2 is submatrix in X. As mentioned above, inflation is used to reflect the ability of replication robustness, and is expected to 1. Bounded distortion reflects the relative value consistency compared to the volume, ensuring that the robust volume guarantees the propositions in Sec. 2. The expected value of bounded distortion is 1.

Proposition 8. *For $\alpha \in (0,1)$, the ClusterRV's inflation has the following inequality: $(1-\alpha)^K \leq inflation \leq (1-\alpha)^{-K}$, where K is the number of clusters.*

Proof 7. ***(Proposition 8)*** *Considering a replication-involving matrix $X_{rep} = replicate(X, c)$, the inflation is written as:*

$$\begin{aligned} inflation &= \frac{ClusterRV(X_{rep})}{ClusterRV(X)} \\ &= \frac{Vol(\widetilde{X}_{rep}) \prod_{i\in K} \rho_{rep,i}}{Vol(\widetilde{X}) \prod_{i\in K} \rho_i} = \frac{\prod_{i\in K} \rho_{rep,i}}{\prod_{i\in K} \rho_i}. \end{aligned} \tag{36}$$

Due to direct copying, the clusters in X_{rep} and X are with similar centers, thus, $Vol(\widetilde{X}_{rep}) \approx Vol(\widetilde{X})$.

Following summation formula of geometric progression, when $0 < \alpha < 1$,

$$1 \leq \rho_{rep,i} := \sum_{p=0}^{\phi_{rep,i}} \alpha^p = \frac{1-\alpha^{(\phi_{rep,i}+1)}}{1-\alpha} \leq \frac{1}{1-\alpha}, \tag{37}$$

$$1 \leq \rho_i := \sum_{p=0}^{\phi_i} \alpha^p = \frac{1-\alpha^{(\phi_i+1)}}{1-\alpha} \leq \frac{1}{1-\alpha}. \tag{38}$$

Thus,

$$(1-\alpha)^K \leq inflation \leq (1-\alpha)^{-K}. \tag{39}$$

The ideal inflation of a data valuation method is 1 meaning complete replication robustness. Reducing $\alpha \to 0$ or K achieves an upper bound on inflation, corresponding to better robustness. However, if α is too small, *ClusterRV* fails to effectively reflect the data points' diversity. This case results in an undesirable effect on data valuation, such as $ClusterRV(X) < Vol(X)$ for a dataset X without replication. If K is too small, the original dataset is compressed to a few clusters. In this case, the *ClusterRV* fails to represent the cluster diversity, causing the data value to be vastly underestimated and resulting in $ClusterRV(X) << Vol(X)$.

Proposition 9. *Let $\alpha = 1/(\beta N)$, where N is the size of dataset. If $N \to \infty$, for any cluster number K, the inflation of ClusterRV will converge to* 1.

Proof 8. ***(Proposition 9)*** *When $\alpha = 1/\beta N$, the following inequality relation exists:*

$$1 \leq \rho_{rep,i} = \sum_{p=0}^{\phi_{rep,i}} \frac{1}{\beta N}^p \leq \frac{1}{1-\frac{1}{\beta N}} = 1 + \frac{1}{\beta N - 1}, \tag{40}$$

$$1 \leq \rho_i = \sum_{p=0}^{\phi_i} \frac{1}{\beta N}^p \leq \frac{1}{1-\frac{1}{\beta N}} = 1 + \frac{1}{\beta N - 1}. \tag{41}$$

Combined with Equation 36,

$$(1 + \frac{1}{\beta N - 1})^{-K} \leq inflation \leq (1 + \frac{1}{\beta N - 1})^{K}. \tag{42}$$

When $N \to \infty$ and K is pre-determined, the following equation exists,

$$\lim_{N\to\infty} (1 + \frac{1}{\beta N - 1})^{-K} = \lim_{N\to\infty} (1 + \frac{1}{\beta N - 1})^{K} = 1. \tag{43}$$

Thus, $inflation \to 1$ is hold under $N \to \infty$.

Proposition 9 theoretically verifies that our proposed *ClusterRV* converges to optimal replication robustness when the datasize N tends to infinity, which is better than **Method 1** with a proved inflation$\to exp(\beta^{-1})$ when $N \to \infty$ [20].

Proposition 10. *Considering $\alpha = 1/(\beta N)$ and $\beta > 0$. For any fixed cluster number K, if the size of dataset $N \to \infty$, the bounded distortion of ClusterRV will converge to* 1.

Proof 9. ***(Proposition 10)*** *When $N \to \infty$, combined with Proposition 9 has proved, the bounded distortion can be rewrittend as:*

$$\begin{aligned} bounded\_distortion &= \frac{ClusterRV(replicate(X_1, c))Vol(X_2)}{Vol(X_1)ClusterRV(replicate(X_2, c))} \\ &\approx \frac{Vol(\widetilde{X_1})Vol(X_2)}{Vol(\widetilde{X_2})Vol(X_1)}. \end{aligned} \tag{44}$$

Taking the scenario where the dataset $X_1 \in R^{N\times M}$ is clustered to K clusters as an example. The matrix of cluster centers is denoted as $\widetilde{X_1} \in R^{K\times M}$. When $N \to \infty$, the number of data points within each cluster are also infinite. Denoting each clusters $C_i, i = 1, ..., K$ contains D data points evenly, the volume of X_1 is as follows.

$$\begin{aligned} Vol(X_1)^2 &= \left| \left[(\widetilde{X_1} + \Gamma_1)^T, ..., (\widetilde{X_1} + \Gamma_D)^T \right] \begin{bmatrix} \widetilde{X_1} + \Gamma_1 \\ \vdots \\ \widetilde{X_1} + \Gamma_D \end{bmatrix} \right| \\ &= \left| \sum_{i=1}^{D} (\widetilde{X_1} + \Gamma_i)^T (X_1 + \Gamma_i) \right|, \end{aligned} \tag{45}$$

where $\Gamma_i = \begin{bmatrix} C_{0,i} - \widetilde{X}_0 \\ \vdots \\ C_{K,i} - \widetilde{X}_K \end{bmatrix} \in R^{K\times M}$ denotes the relative distance of i-th data points in K clusters to theirs cluster's centers. Each cluster center generated by k-Means is:

$$\widetilde{X}_k = \frac{\sum_{i=1}^{D} \delta_{ik} x_i}{\sum_{i=1}^{D} \delta_{ik}}, \tag{46}$$

where δ_{ik} is a cluster indicator variable with $\delta_{ik} = 1$ if x_i in k-th cluster. As the number of data points increases, the distance vectors of the points to cluster center are cancel each other out,

$$\sum_{i=0}^{D} \Gamma_i \to \overrightarrow{0}. \tag{47}$$

In the same way,

$$Vol(X_2)^2 = \left| \sum_{i=1}^{D} (\widetilde{X}_2)^T (\widetilde{X}_2) + \sum_{i=1}^{D} (\widetilde{\Upsilon}_i)^T (\Upsilon_i) \right|. \tag{48}$$

Applying determinant property,

$$\begin{aligned} \frac{Vol(X_1)^2}{Vol(X_2)^2} &= \frac{\left| D \cdot (\widetilde{X}_1)^T (\widetilde{X}_1) + \sum_{i=1}^{D} (\Gamma_i)^T (\Gamma_i) \right|}{\left| D \cdot (\widetilde{X}_2)^T (\widetilde{X}_2) + \sum_{i=1}^{D} (\Upsilon_i)^T (\Upsilon_i) \right|} \\ &= \frac{\left| (\widetilde{X}_1)^T (\widetilde{X}_1) + \frac{1}{D} \sum_{i=1}^{D} (\Gamma_i)^T (\Gamma_i) \right|}{\left| (\widetilde{X}_2)^T (\widetilde{X}_2) + \frac{1}{D} \sum_{i=1}^{D} (\Upsilon_i)^T (\Upsilon_i) \right|}. \end{aligned} \tag{49}$$

Considering $\Gamma_i^T \Gamma_i = \alpha_i \widetilde{X}_1^T \widetilde{X}_1$, for the data size $N \to \infty$, there will also exists Υ_i in infinite space satisfied $\Upsilon_i^T \Upsilon_i = \alpha_i \widetilde{X}_2^T \widetilde{X}_2$. Whether $\Gamma_{i,k}$

and $\Upsilon_{i,k}$ *belong to same clusters* k *is not necessary, because:*

$$\begin{aligned}\Gamma_i^T\Gamma &= \begin{bmatrix}\Gamma_{i,1}^T & \dots & \Gamma_{i,K}^T\end{bmatrix}\begin{bmatrix}\Gamma_{i,1}\\ \dots \\ \Gamma_{i,K}\end{bmatrix} \\ &= \sum_{k=1}^{K}\Gamma_{i,k}^T\Gamma_{i,k} = \underbrace{\begin{bmatrix}\Gamma_{i,K}, \dots & \Gamma_{i,1}\end{bmatrix}}_{any\ order}\begin{bmatrix}\Gamma_{i,K}\\ \dots \\ \Gamma_{i,1}\end{bmatrix}.\end{aligned} \tag{50}$$

Then, the Equation 49 is rewritten as:

$$\begin{aligned}\frac{Vol(X_1)^2}{Vol(X_2)^2} &= \frac{\left|(\widetilde{X}_1)^T(\widetilde{X}_1) + \frac{1}{D}\sum_{i=1}^{D}(\Gamma_i)^T(\Gamma_i)\right|}{\left|(\widetilde{X}_2)^T(\widetilde{X}_2) + \frac{1}{D}\sum_{i=1}^{D}(\Upsilon_i)^T(\Upsilon_i)\right|} \\ &= \frac{\left|I + \frac{1}{D}\sum_{i=1}^{D}\alpha_i\right|}{\left|I + \frac{1}{D}\sum_{i=1}^{D}\alpha_i\right|}\frac{\left|(\widetilde{X}_1)^T(\widetilde{X}_1)\right|}{\left|(\widetilde{X}_2)^T(\widetilde{X}_2)\right|} = \frac{Vol(\widetilde{X}_1)^2}{Vol(\widetilde{X}_2)^2}.\end{aligned} \tag{51}$$

Thus, when $N \to \infty$*, the data points in each cluster* $D \to \infty$*,*

$$bounded\_distortion \approx \frac{Vol(\widetilde{X_1})Vol(X_2)}{Vol(\widetilde{X_2})Vol(X_1)} \to 1. \tag{52}$$

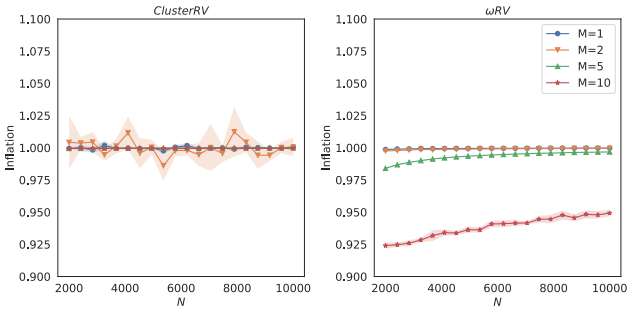

Figure 6: *ClusterRV* **(left) vs.** ωRV **(right) in the inflation metric under replication dataset** $X_{ref} = replicate(X, 10)$**. Solid lines represent statistical mean and shadow represents statistical standard error under** 100 **trials, respectively.**

In terms of $inflation$ and $bounded\ distortion$, we experimentally compare the replication robustness of our proposed *ClusterRV* to ωRV in **Method 1** [20]. To ensure a fair comparison, we compress the data set $X \sim \mathcal{N}(0,1)^M$ to the same dimensionality, with the paramater settings are $\omega = 0.5$ in ωRV, $K = 2^M$ in *ClusterRV*. The diversity parameter β in both methods is set to 0.1. All experiments are performed 100 trials to ensure statistical significance. Figures 6 shows performance of *ClusterRV* and ωRV on inflation, where the valued dataset X_{rep} is replicated from X for $c = 10$ times. *ClusterRV* shows better inflation performance for the mean value is closer to 1 especially under high-dimension, which is in line with **Proposition 9**. For clustering randomness in each trial, *ClusterRV* is with certain but acceptable variance in some trials, this issue is improved in higher feature dimensions. Figure 7 and Figure 8 show performance of *ClusterRV* and ωRV on bounded distortion under non-replication dataset and replicated dataset, respectively.

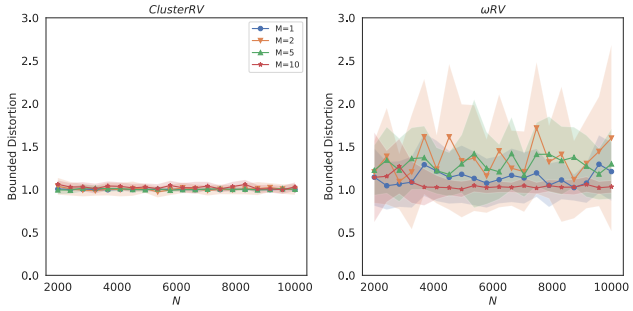

Figure 7: *ClusterRV* **(left) vs.** ωRV **(right) in bounded distortion under non-replication dataset** $X \sim \mathcal{N}(0,1)^M$**. Solid lines represent statistical mean and shadow represents statistical standard error under** 100 **trials, respectively.**

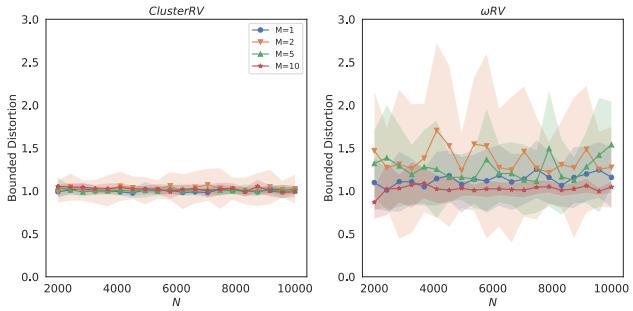

Figure 8: *ClusterRV* **(left) vs.** ωRV **(right) in bounded distortion under replication dataset** $X_{ref} = replicate(X, 10)$**. Solid lines represent statistical mean and shadow represents statistical standard error under** 100 **trials, respectively.**

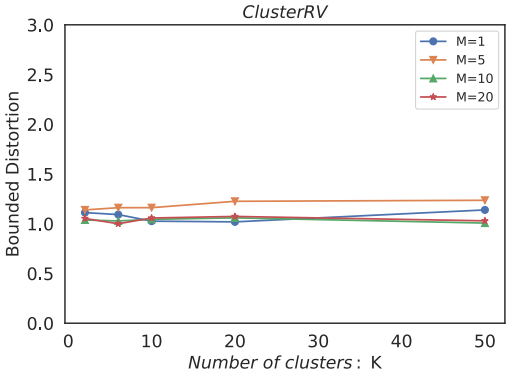

Figure 9: K-sensitivity of ClusterRV.

In both cases, the results show that the *ClusterRV* outperforms ωRV statistically with the bounded distortion closer to 1.

In Figure 9, we conduct experiment to visualize the parameter-sensitivity of *ClusterRV* to cluster number K. In the expirement settings, the subset $X1, X2 \in R^{5000\times M}$ and randomly sampled from

$N(0,1)$, where $M \in [1,5,10,20]$. The result shows that under different number of clusters: $K \in [2,6,10,20,50]$, the *ClusterRV* presents lower parameter-sensitivity with limited fluctuation. This insensitivity to K selection is more obvious under high-dimensional features. The experimental results provide a guidance for selecting the initial number of clusters, when the computational burden for linear search is too high to perform. An K following the feature dimension is better.

We conclude this section by a *ClusterRV*-based Shapley value (*ClusterRVSV*), which is suitable for data valuation on subsets with replication and outliers injection in SsM data market. For any subset X_S in X, the *ClusterRVSV* is defined in Equation 53. Instead of requiring the data vendor or market to perform extensive preprocessing work to detect and remove malicious data, data valuation based on cluster RVSV enables the data market to perform direct and reasonable pricing on datasets containing replication and outliers.

$$\psi_{ClusterRVSV}(X_S) = \sum_{X_T \subset \{X\} \setminus \{X_S\}} \frac{|X_T|!(|X|-|X_T|-1)!(ClusterRV(X_T \cup X_S) - ClusterRV(X_T))}{|X|!}. \tag{53}$$

4 REVENUE OPTIMIZATION IN SVBP

In the aforementioned sections, we have introduced the *ClusterRVSV* to offer subsets' data value for pricing in the SsM market while enhancing market robustness under replication and outliers. In this section, we will study how to design SVBP to maximize the vendor's revenue and satisfy arbitrage-free in SsM data market. In particular, SVBP should be satisfied the following requirements:

- Non-negativity: A pricing function p should be non-negative for any dataset with value v: $p(v) \geq 0$;
- Fairness: For a data market to work well, datasets with lower data value should not be assigned higher prices. A pricing function should have a positive correlation with data value, that is: if $v_1 \geq v_2$, $p(v_1) \geq p(v_2)$;
- Arbitrage-free: The arbitrage is that buyers can pay for sub-datasets separately with a lower price than paying for the whole dataset in once. In a healthy data market, arbitrage should be avoided. Considering sub-datasets with data value v_1 and v_2, respectively. An arbitrage-free pricing function should follow: $p(v_1 + v_2) \leq p(v_1) + p(v_2)$.

Rather than dealing with a continuous pricing function $p(v)$ directly, we samples n discrete points of the form $(v_i, p(v_i))$ for programming solvability. The buyers who are interested in subscribing a sub-dataset with value v_i have a willingness-to-pay (WTP) value $w_i = w(v_i)$, where WTP function w refers to upper bound price at which a consumer is willing to subscribe the data. The buyers will buy the subset only if $p(v_i) \leq w(v_i)$. Considering the demand distribution function $d(v)$ denoting how many buyers are interested in a dataset with value v, the profit of the data vendor setting the price at $p(v)$ is $\mathcal{PF} = d(v)p(v) \cdot \mathbf{1}_{p(v) \leq w(v)}$, where $\mathbf{1}_{p(v) \leq w(v)}$ is an indicator variable that takes value 1 if $p(v) \leq w(v)$, otherwise 0. Formally, we desire SVBP to be a feasible solution to the following revenue maximization problem while satisfying the aforementioned requirements:

$$\begin{aligned} \max_{p^*} f(p(v_1), ..., p(v_n)) &= \sum_{i=1}^{n} d_i p(v_i) \cdot \mathbf{1}_{p(v_i) \leq w(v_i)} \\ subject\ to \quad & p(v_i + v_j) \leq p(v_i) + p(v_j), \\ & p(v_j) \geq p(v_i), v_j \geq v_i, \\ & p(v_i) \geq 0, v_i \geq 0, \\ & i, j \in \{1, ..., n\} \ \ and \ i \neq j. \end{aligned} \tag{54}$$

The arbitrage-free and fairness requirements lead to sub-additivity and monotone constraints in Equation 54. The third constraint corresponds to the non-negativity requirement. However, the proof of the existence of linear interpolation on the solution in Equation 54 with sub-additivity and monotone constraint is coNP-hard.

Definition 4. ***(Unbounded subset-sum problem)*** *Given a set of positive integers $\{k_0, ..., k_n\}$, an unbounded subset-sum problem is defined as to find the non-negative integers α_i so that $\sum_{i=1}^{n} \alpha_i k_i = K$, for we can achieve K by k_i for any times, it's known that finding the solution of unbounded subset-sum problem is NP-hard.*

Lemma 2. *Let $v_i = p_i, i = 1, ..., n$, $p_{n+1} = K + \triangle$ when $v_{n+1} = K$ and $\triangle \in (0,1)$, a subadditive and monotone function $p(x)$ interpolating on the points (v_i, p_i) exist if and only if unbounded subset sum $\sum_{i=1}^{n} \alpha_i v_i = K$ not exists.*

Proof 10. ***(Lemma 2)*** *If $\sum_{i=1}^{n} \alpha_i v_i = K$ exists, then we have:*

$$K + \triangle = p(K) = p(\sum_{i=1}^{n} \alpha_i v_i) \leq \sum_{i=1}^{n} \alpha_i p_i \overset{v_i = p_i}{=} K. \tag{55}$$

$K + \triangle = K$ is a contradiction so that if $\sum_{i=1}^{n} \alpha_i v_i = K$ exists, a subadditive and monotone function $p(x)$ interpolating on the $n+1$ points (v_i, p_i) is not exist.

Conversely, in the next, we prove that if $\sum_{i=1}^{n} \alpha_i v_i = K$ not exists, a subadditive and monotone function $p(x)$ that interpolates the $(n+1)$ points can be constructed. First, we introduce a function $\mu(x)$ to reflect the smallest possible unbounded subset sum at x. $\mu(x)$ at least contrains an unbounded subset sum constains x, thus $\mu(x) \geq x$. Then, we define a function $p(x) = min(\mu(x), K + \triangle)$ and our goal is to prove such $p(x)$ is satisfied subadditive, monotone and interpolating on the $n+1$ points (v_i, p_i). It is apparent that $p(x)$ is monotone. Since a set containing x-self is a minumum unbounded subset sum, we have $\mu_i = v_i = p_i \leq K + \triangle$. For we have assumed that $\sum_{i=1}^{n} \alpha_i v_i = \sum_{i=1}^{n} \alpha_i p_i = K$ is not exist, thus $\mu(v_{i+1}) \geq K + 1$. Then the $p(x)$ can be written as:

$$p(x) = \begin{cases} \mu(x), & \mu(x) \leq K \\ K + \triangle, \mu(x) \geq K + 1 \end{cases} \tag{56}$$

If $\mu(x) \geq K+1$, then $p(x+y) \leq K+\triangle = p(x) \leq p(x)+p(y)$. When both $\mu(x) \leq K$ and $\mu(y) \leq K$, we have $p(x) = \mu(x) = \sum_{i=1}^{n} \alpha_i v_i$ and $p(y) = \mu(y) = \sum_{i=1}^{n} \beta_i v_i$. Then, $x+y \leq p(x)+p(y) = \sum_{i=1}^{n} (\alpha_i + \beta_i) v_i$. According to the definition of $\mu(x)$, $p(x+y) = \mu(x+y) = min(x+y, \sum_{i=1}^{n} \gamma_i v_i) \leq \sum_{i=1}^{n} (\alpha_i + \beta i) v_i = p(x) + p(y)$. Above all, we have proved Lemma 2. For the unbounded subset-sum problem is NP-hard, proving the unbounded subset sum $\sum_{i=1}^{n} \alpha_i v_i = K$ not exists in Lemma 2 is a co-NP hard problem.

A solution is to relax the a sub-additivity constraint to a approximate version: $p(x)/x \geq p(y)/y$ for every $x \leq y$. The reformulated

optimization problem is as follows:

$$\begin{aligned} &\max_{p^*} f(p(v_1), ..., p(v_n)) = \sum_{i=1}^{n} d_i p(v_i) \cdot \mathbf{1}_{p(v_i) \le w(v_i)} \\ &subject\ to \quad p(v_i)/v_i \ge p(v_j)/v_j, v_j \ge v_i, \\ &\qquad\qquad p(v_j) \ge p(v_i), v_j \ge v_i, \\ &\qquad\qquad p(v_i) \ge 0, v_i \ge 0, \\ &\qquad\qquad i, j \in \{1, ..., n\} \ \ and \ i \ne j. \end{aligned} \tag{57}$$

Proposition 11. *Given a pricing function p satisfying the constraint of $p(x)/x \ge p(y)/y, x \le y$, it also satisfy $p(x+y) \le p(x)+p(y)$ strictly.*

Proof 11. ***(Proposition 11)*** *For the pricing function p satisfies $p(x)/x \ge p(y)/y$ when $x \le y$, there exists:*

$$\begin{aligned} &\frac{p(x+y)}{x+y} \le \min\left(\frac{p(x)}{x}, \frac{p(y)}{y}\right) \Rightarrow \\ &p(x+y) \le \min\left(p(x) + \underbrace{\frac{yp(x)}{x}}_{\ge p(y)}, p(y) + \underbrace{\frac{xp(y)}{y}}_{\ge p(x)}\right) \\ &\le p(x) + p(y). \end{aligned} \tag{58}$$

Constraint $p(x)/x \ge p(y)/y, x \le y$ representing a subspace of sub-additivity constraint is proved.

Since constraint $p(x)/x \ge p(y)/y, x \le y$ represents a subspace in the original sub-additivity constraints, the optimal solution of Equation 57 is the feasible solution of Equation 54. Denoting the sampled points from pricing function as discrete variables z_i, the Equation 57 can be equivalently rewritten as a optimization problem in discrete space and the pricing function can be obtained by interpolation on z_i. The discrete variables optimization problem is as follows:

$$\begin{aligned} &\max_{z} f(z_1, ..., z_n) = \sum_{i=1}^{n} d_i z_i \cdot \mathbf{1}_{z_i \le w_i} \\ &subject\ to \quad z_j/v_j \le z_i/v_i, v_j \ge v_i, \\ &\qquad\qquad z_j \ge z_i, v_j \ge v_i, \\ &\qquad\qquad z_j \ge 0, \\ &\qquad\qquad i, j \in \{1, ..., n\} \ \ and \ i \ne j. \end{aligned} \tag{59}$$

5 EXPERIMENTS AND DISCUSSION

In this section, we first verify **Corollary 2** in Sec. 2 that a larger volume leads to a larger parameter variation, and derive some practical purposes in Sec. 5.1. Subsequently, in Sec. 5.2, we show that the proposed *ClusterRVSV* produces consistent value-allocation in multi-tasks, and we additionally demonstrate the limitations of existing other methods. Then, in Sec. 5.3, we verify the robustness guarantee of *ClusterRVSV* in replication and outliers. Overall, only our proposed *ClusterRVSV* can satisfy task-agnostic, replication and outliers robustness compared to other baselines. The pricing strategy, SVBP, in the subset-sale mode data market is viable based on the findings of the experiments above. Lastly, we propose a case study of the subset-sale model data market in Sec. 5.4. Importantly, we validate that the proposed pricing strategy, SVBP, generates more revenue for the data vendors while guaranteeing higher affordability for buyers. All experiments were run on a server with Intel(R) Xeon(R) @ 2.20GHz processor and 128GB RAM.

5.1 A Larger Volume Leads To A Larger Parameter Variation

In this subsection, we consider the paradigm of sequentially training a new model on each subset ordered by volume to observe the model's parameter variation trend. The subsets are ordered by three sorting methods, which are volume from highest to lowest (blue bars in the Figure 10, 11), volume from lowest to highest (orange bars in the Figure 10, 11), and random (green bars in the Figure 10, 11), respectively. In order to make our results more generalizable, we conduct experiments under a regression NN model (Figure 10) and a classification NN model (Figure 11), respectively. The regression NN is a single-layer perceptron with MSE loss function, and the classification NN is a 3-layers perceptron with cross-entropy loss. We use a real-world house sale prices dataset [29] for the regression model and a real-world breast cancer diagnosis dataset [30] for the classification model. Both two datasets have been pre-processed to contain 5 standardized features, respectively. To verify that the **Corollary 2** holds under different forms of subsets, each dataset is divided into row subsets,$\{X_1, X_2, X_3, X_4, X_5\}, X_i \in R^{100\times 5}$, and column subsets $\{X_1, X_2, X_3, X_4, X_5\}, X_i \in R^{500\times 1}$, respectively. The left graphs in Figure 10, 11 depict the NN models' parameter variation trends when training on row subsets of varying volume. The right graphs in Figure 10, 11 depict the NN models' parameter variation trends when training column subsets of varying volume. The volume in this subsection is calculated by the proposed *ClusterRV*. All experiments are performed 100 times to ensure statistical significance.

We observe the trend that training a model on a subset with a higher (resp., lower) volume value leads to a higher (resp., lower) model parameter variation, thus verifying the **Corollary 2**. Furthermore, this trend is observed in subsets of various forms, multi-model structures, and multi-loss functions, verifying that the **Corollary 2** holds in generalized scenarios.

The above findings give a subset valuation solution for SsM. Furthermore, because the volume is proportional to the model parameter variation and dataset size, it provides a more mutually satisfactory value for both the vendor and the buyer. In the next, we verify that the volume-based data value provides guidance for buyers on a limited budget to purchase subsets preferentially in the data market (Figure 1). We employ the same model selection and dataset selection as above. Each subset $X_i \in R^{500\times 1}, i = 1, 2, 3, 4, 5$ contains a unique feature and is designed to simulate a market scenario in which a buyer with a limited budget does not know how to select the feature dimension. We consider the paradigm of sequentially adding subsets X_i into the training set X_{train} for training the model. At the same time, we observe the model loss on validation set X_{val} to reflect the trend in model performance. The validation set X_{val} contains the same feature dimensions as the training set but different data points with 100. In the regression model (left graph in Figure 12), we find that when trained on a set containing the subsets $1 \sim 3$ ordered by the highest value, the model's validation loss is close to one trained on the whole subset.

In the classification model (right graph in Figure 12), a training set containing subsets $1 \sim 4$ ordered by the highest value is enough for near-optimal model performance. The result verifies that buyers can achieve better model performance (lower validation loss) with fewer subsets by prioritizing data subsets with higher value first.

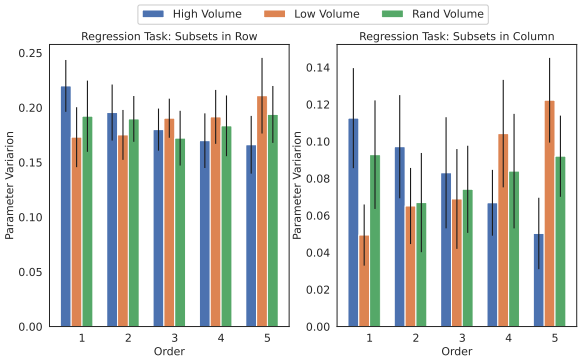

Figure 10: The parameter variation of a regression NN model trained on each row subset (left) and column subset (right) sorted by volume. All experiments are performed 100 times to ensure statistical significance.

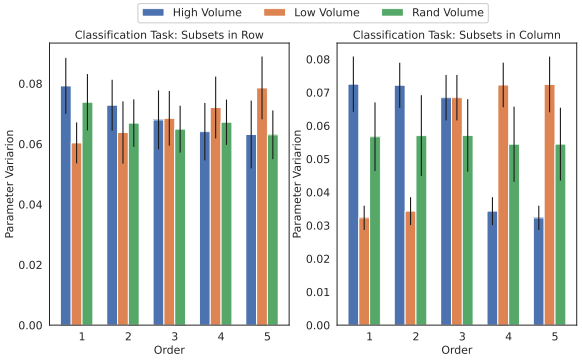

Figure 11: The parameter variation of a classification NN model trained on each row subset (left) and column subset (right) sorted by volume. All experiments are performed 100 times to ensure statistical significance.

5.2 Data Valuation in Multi-Tasks

In order to verify task-agnostic, we compare the subses' value calculated by volume-based methods and other validation-based baselines in different task scenarios. In this experiment, we use a real-word dataset *WineQuality* [31]. The dataset is pre-processed to 10 feature columns and further to be standardized. For diversity , we divide the dataset into 4 subsets as follows: X_1 {"pH" ≤ 3.1} $\subset$ X_2 {"pH" ≤ 3.2}, X_4 {"pH" > 3.4} $\subset$ X_3 {"pH" > 3.2}, X_2 {"pH" ≤ 3.2} $\cap$ X_3 {"pH" > 3.2} $= \varnothing$. For observing easily, the summation of the subsets' value is normalized to 1 and the proportion of each subset is shown in the Figure 13. Each subset is

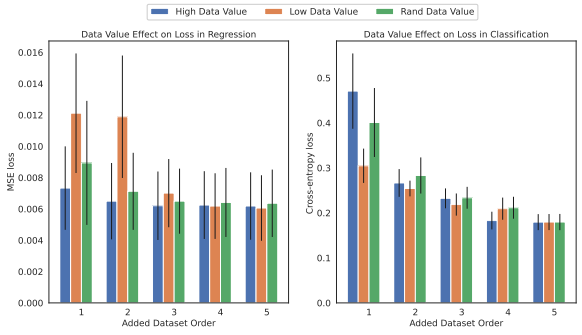

Figure 12: The validation losses of regression NN model (left) and classification NN model (right) trained on subsequently added feature subsets sorted by volume. All experiments are performed 100 times to ensure statistical significance.

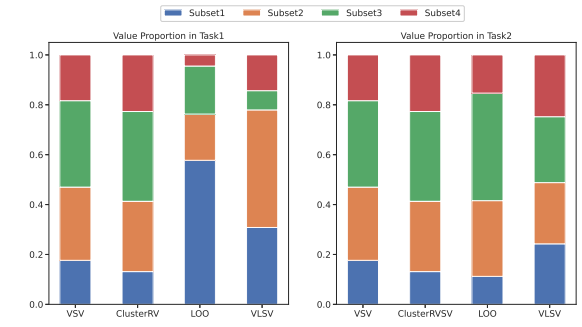

Figure 13: The value proportions of four subsets in different tasks, pH regression task (left) and quality classification task (right).

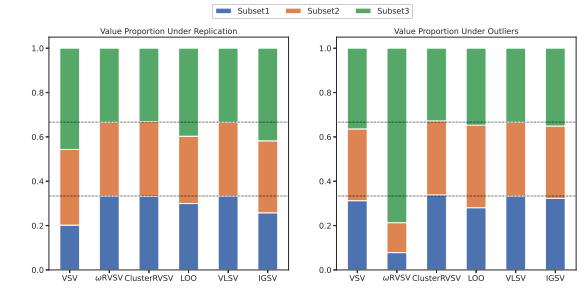

Figure 14: The value proportions of three subsets under two scenarios: dishonest replication (left) and outliers injection (right).

used in two downstream tasks, including a pH regression task (left graph in Figure 13) and a *quality* classification task (right graph in Figure 13). The regression NN model is a 3-layers perceptron with MSE loss function, and the classification NN model is a 3-layers perceptron with cross-entropy loss.

As shown in Figure 13, the subset value calculated by VSV and $ClusterRVSV$ respectively is consistent in different tasks, thus, is task-agnostic. In contrast, the subset value calculated by LOO and $VLSL$ varies significantly in different tasks, thus, is task-dependent. Pricing based on the data value calculated by task-dependent method results in price discrimination across tasks. Although market segmentation operation has been used to design different prices for buyers with different characteristics and payment levels, no research has proven the feasibility of assigning a dataset with different prices according to different tasks for the same buyer. Thus, avoiding discrimination among various tasks should still be an essential criterion for data valuation. Our experimental results show that volume-based data valuation can avoid the problem of price discrimination across tasks.

5.3 Data Valuation in Replication and Outliers Injection

In this experiment, we simulate replication and outlier injection scenarios by constructing the subsets as following: 1) replication-injection subsets $\{X_1, X_2, X_3\}$, where X_1 is a non-replicated subset contains 100 data points, X_2 and X_3 are the replicated subsets generated by copying X_1 for 2 and 4 times, respectively; 2) outliers-injection subsets $\{X_1, X_2, X_3\}$, where X_1 contains 100 data points. The $X_{outlier1}$ containing single outlier and $X_{outlier2}$ containing 5 outliers are sampled from the invalid data points of $WineQuality$ by $Z\text{-}score > 3$, thus, $X_2 = \{X_1, X_{outlier1}\}$ is with 1% outlier ratio, and $X_3 = \{X_1, X_{outlier2}\}$ is with 5% outlier ratio. For comparison, we introduce more validation-free baselines, including the ω-$based$ robust volume Shapley value ($\omega RVSV$) in **Method 1** [20], the information gain Shapley value ($IGSV$) [32]. For observing easily, the summation of the subsets' value is normalized to 1 and the proportion of each subset is shown in the Figure 14. We desire a dataset value under replicate-injection or outlier-injection to be consistent with the value of a clean dataset. The left graph in Figure 14 shows the subset value proportion under replication. In this expirement, The hyparameter ω in $\omega RVSV$ is set to 0.1 and the clusters number K in $ClusterRVSV$ is set to 4 according to the dataset label "quality" mainly containing 4 classes given in the metadata. When calculated by $\omega RVSV$, $ClusterRVSV$, and $VLSV$, the data value of subsets X_1, X_2, X_3 are the same. That is, when introduced replication, X_2 and X_3 do not generate a higher value than X_1. Therefore, these three methods are replication-robust, but differ in the mechanism for achieving robustness. $\omega RVSV$ and $ClusterRVSV$ obtain replication robustness by data compression, and $VLSV$ obtains replication robustness for replication data sets fail to increase models' validation loss. In Sec. 3, we demonstrate the difference in data compression between $\omega RVSV$ and our proposed $ClusterRVSV$, and the subsequent experiments verify the limitation of $\omega RVSV$ in outlier-injection. The left graph in Figure 14 also shows a noticeable increases of X_2/X_3's value proportion in methods of LOO and $IGSV$, implying that LOO and $IGSV$ are not robust under replication.

The right graph in Figure 14 shows the subset value proportion under outliers.The hyparameter ω in $\omega RVSV$ is set to 0.5 to reduce the outlier-sensitivity. The clusters number K in $ClusterRVSV$ is still set to 4 as aforementioned. When calculated by $ClusterRVSV$, and $VLSV$, the data value of subsets X_1, X_2, X_3 are the same. That is, when introduced outliers, X_2 and X_3 do not generate a higher value than X_1. Therefore, $ClusterRVSV$, and $VLSV$ are outlier-robust. Their difference is that $ClusterRVSV$ counteracts the effect of outliers by clustering, while $VLSV$ obtains replication robustness for replication data sets fail to increase models' validation loss. In VSV, $\omega RVSV$, LOO, $IGSV$, there exists a increases of X_2 and X_3's value proportion, implying these methods are not robust under outliers. Especially in $\omega RVSV$, the value of X_3 far exceeds the value of a clean subset X_1. The discretization mechanism in $\omega RVSV$ fails to eliminate outliers from cubes and even magnify its effect in data value after dataset compression.

Overall, when compared to existing baselines, $ClusterRVSV$ is the only data valuation method that can apply to subsets in various forms, with task-agnostic, replication-robust, and outlier-robust properties.

5.4 Revenue Maximization in Subset-sale Mode Data Markets

The preceding sections and experiments verify that $ClusterRVSV$ meets the properties (Sec. 1) expected of data valuation in the SsM data market. In this subsection, we demonstrate the advantages of our proposed pricing function, SVBP, on the vendor's revenue and buyer's affordability ratio (fraction of the buyers that can afford to buy/subscribe instance they require) compared to a FsM data market with flat pricing strategies. Two widely used pricing strategies in FsM are considered as baselines [33].

- **MaxW** assigns a flat price to all instances, based on the highest value in the buyer's WTP curve.
- **MedW** assigns a flat price to all instances, ensuring that at least half of the buyers can afford their desired instance.

We use sequential (least-squares) quadratic programming (SQP) algorithm to find the feasible solution of the maximization problem in Equation 59. The initial points of SQP are sampled from the linear pricing function: $p(v) = a \times v$, $a = 0.5$. The experimental results are shown in Figure 15.

For generality, we simulated a variety of market scenarios. One scenario is that the demand distributions are fixed, but the WTPs are in various forms, as shown in Figure 15 (a), (b) and (c). The top sub-figures give the demand distributions following $bimodal$ and WTP curves in $concave$, $convex$ and $logistic$ forms, respectively. The middle graphs give the price functions under SVBP, MaxW, and MedW strategy, respectively. The bottom graphs give the vendor's revenue and buyer's affordability ratio of SVBP compared to other baselines. The results show that SVBP achieves at least 23.29× and 1.49× revenue gain compared to the MaxW and MedW, under various WTPs. The affordability ratio of SVBP is close to 100% when WTP is in $concave$ form and is 50% when WTP is in $convex$ or $logistic$ form. In a market, the max revenue with 100% affordability happens when pricing function is consistent with WTP curve. When the WTP is convex with an increasing slope, a pricing function consistent with WTP violates the arbitrage-free constraint for $z_j/v_j \geq z_i/v_i$ happening at $v_j \geq v_i$. To avoid this case, a trade-off of SVBP is to increase the slope of the price function curve for low-value data. That's why the affordability ratio under $SVBP$ is 50% when WTP is in $convex$ or $logistic$ form.

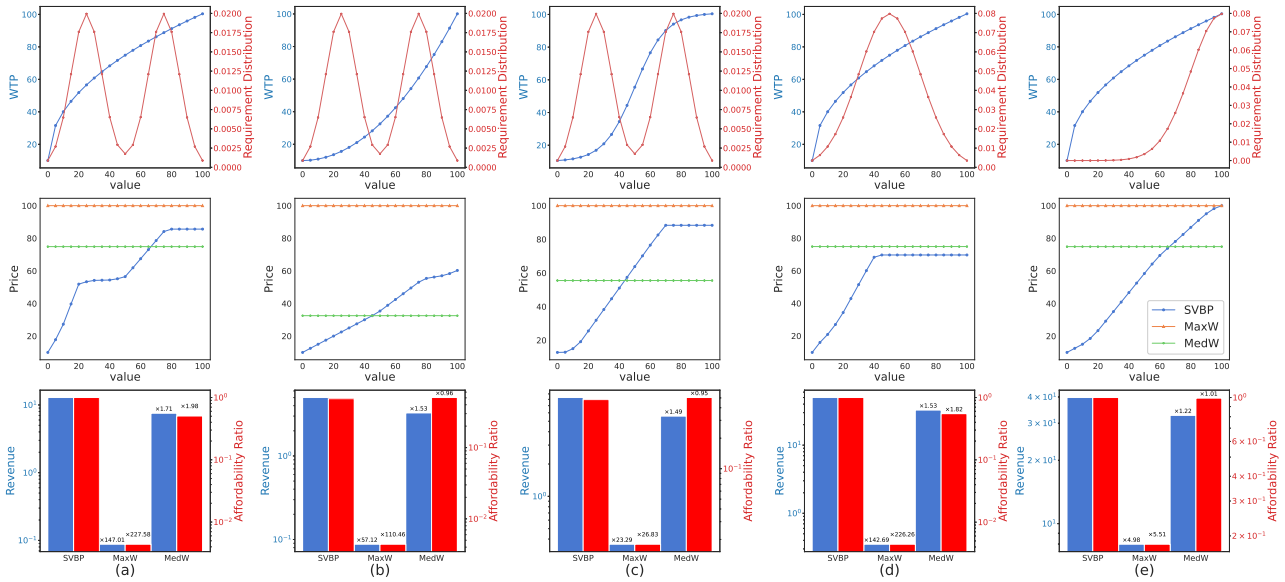

Figure 15: The pricing functions and their corresponding revenue and affordability ratio under different WTP curves and demand distribution curves. The demand distribution curves in figure (a), (b), (c) follow a bi-modal distribution of $\frac{\mathcal{N}(25,10)+\mathcal{N}(75,10)}{2}$. The WTPs in figure (a), (d), (e) are in concave form of $w = 10v^{0.478} + 10$. The WTPs in figure (b) is in convex form of $w = v^2/111 + 10$. The WTP curve in figure (c) is in logistic form of $w = 91/(1 + e^{-(0.1(v-50))}) + 10$. The demand distribution in figure (d) follows a normal distribution of $\mathcal{N}(50, 20)$ and the demand distribution in figure (e) follows a normal distribution of $\mathcal{N}(100, 20)$.

The other scenario is that the WTP curves of buyers are fixed, but the demands follow various distributions, as shown in Figure 15 (a), (d) and (e). The results show that SVBP can change price adaptively according to the changing of demand distributions. When the majority of demands concentrate on datasets with high value, the SVBP tends to produce a price function that its upper bound ties close to the price of high-value datasets (Figure 15 (e)). When the majority demands concentrate on datasets with medium value, the upper bound price of SVBP tends to the price of dataset with medium value (Figure 15 (d)). When the demand follows $bi-modal$ distribution, SVBP shows a stepped pricing manner (Figure 15 (a)). For adaptively changing the price under different demands, SVBP generates up to 4.98× and 1.22× revenue gain compared to MaxW and MedW, and all affordability ratios under various demand distributions are promoted to 100%.

Overall, a SsM data market, where SVBP is allowed to adjust to different WTPs and distributions , achieves more revenue/affordability than FsM data market with a flat pricing strategy.

6 RELATED WORK

None of previous work focused on investigating the sales of subsets of a dataset for neural networks in a data marketplace. In this section, we only discuss related work on data valuation in a data market, especially for neural networks as downstream models. A mature pricing strategy aims at maximizing the profit rather than reducing the cost [11]. [33] proposed a model-based pricing for data markets directly selling the products of machine learning models. More intuitively, we propose a subset-based pricing strategy, SVBP, without changing the data product to other product form. The most technical challenge in SVBP is an efficient data valuation method on subsets. The most widely used data valuation method is the validation-based Shapley value (VSV) [14, 34]. Yoon et al. proposed a reinforcement learning (RL) based data valuation [35]. The limitation of both VSV and RL lies on the fact that they are highly coupled with a suitable validation set. Thus, they are task-dependent. [20] adopted volume for task-agnostic data valuation but limited in row data (data points) valuation and linear regression tasks [20]. For the robust data valuation under replication and outliers, a similarity metric was used to construct a "robust-to-replication" version of the Shapley value algorithm [36]. Han et al. defined a whole family of replication-robust payoff allocations, including the Banzhaf value and Leave-one-out [37]. These methods also required to carefully select validation sets. The work of [20] discretized a replicated dataset X into a set of cubes by a hyparameter ω and use *mean* for cube-compression to eliminate replicated data, but such method is highly outlier-sensitive. Tuning a suitable ω is also intractable when features are distributed in various scales. For outlier-robustness of data valuation, existing methods demonstrated certain effectiveness using statistical analysis [38] and machine learning [39] to detect and wash out outliers, but imposing computational burden on data vendors and data markets.

7 CONCLUSION

In this paper, we have formalized the subset-sale mode (SsM) data market mechanism, allowing buyers to purchase subsets of a dataset at lower budgets. We have proposed a subset-value base pricing (SVBP) strategy suitable for this novel market mechanism. We have also presented a cluster-based robust volume (*ClusterRV*) as

a value metric for the subset valuation. We then have combined the proposed *ClusterRV* with Shapley value to obtain a novel dataset valuation with the task-agnostic property and robustness under dishonest behaviors from data vendors. Following the accessibility of the data value, we have formalized arbitrage-freeness as a constraint and provided the revenue optimization problem to obtain a feasible solution for SVBP. Extensive experiments have validated that *ClusterRVSV* outperforms other baselines in various scenarios. Our results have demonstrated that the SsM data market with SVBP provides a higher revenue to data vendors and a higher affordability to buyers than those of the current fullset-sale mode (FsM) data market with a flat pricing strategy.

REFERENCES

[1] Chen Li et al. 2022. Illumination angle correction during image acquisition in light-sheet fluorescence microscopy using deep learning. *Biomed. Opt. Express*, 13(2):888–901, 2022.
[2] Chen Li et al. 2021. Deep learning-based autofocus method enhances image quality in light-sheet fluorescence microscopy. *Biomed. Opt. Express*, 12(8):5214–5226, 2021.
[3] Chenhan Zhang et al. 2019. Spatial-temporal graph attention networks: A deep learning approach for traffic forecasting. *IEEE Access*, 7:166246–166256, 2019.
[4] Dawex. https://www.dawex.com/en/.
[5] Worldquant. https://www.dawex.com/en/.
[6] Vaishali Ravindranath et al. 2020. Swarm intelligence based feature selection for intrusion and detection system in cloud infrastructure. In *IEEE CEC*, 2020.
[7] Mohammad Sultan Mahmud et al. 2020. A survey of data partitioning and sampling methods to support big data analysis. *Big Data Min. Analyt.*, 3(2):85–101, 2020.
[8] Chao Li et al. 2014. A theory of pricing private data. *TODS*, 2014.
[9] Hal R Varian. 1989. Price discrimination. *Handbook of industrial organization*, 1989.
[10] Jun Liu and Francis A Longstaff. 2004. Losing money on arbitrage: Optimal dynamic portfolio choice in markets with arbitrage opportunities. *Rev. Financ. Stud.*, pages 611–641, 2004.
[11] Fan Liang et al. 2018. A survey on big data market: Pricing, trading and protection. *IEEE Access*, 6:15132–15154, 2018.
[12] Ruoxi Jia et al. 2020. Towards efficient data valuation based on the shapley value. In *AISTATS*, 2019.
[13] I Elizabeth Kumar et al. 2020. Problems with shapley-value-based explanations as feature importance measures. In *ICML*, 2020.
[14] Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *ICML*, 2019.
[15] Michal Derezinski and Manfred KK Warmuth. 2017. Unbiased estimates for linear regression via volume sampling. In *NIPS*, 2017.
[16] John Wright et al. 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, 2009.
[17] Quoc Le et al. 2011. Ica with reconstruction cost for efficient overcomplete feature learning. In *NIPS*, 2011.
[18] Ruoxi Jia et al. 2021. Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification? In *IEEE/CVF CVPR*, 2021.
[19] Tom Yan and Ariel D Procaccia. 2021. If you like shapley then you'll love the core. In *AAAI*, 2021.
[20] Xinyi Xu et al. 2021. Validation free and replication robust volume-based data valuation. In *NIPS*, 2021.
[21] Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *arXiv:1207.6083*, 2012.
[22] Grigorios Tzortzis and Aristidis Likas. 2014. The minmax k-means clustering algorithm. *Pattern Recognit.*, 47(7):2505–2516, 2014.
[23] Yanchi Liu et al. 2010. Understanding of internal clustering validation measures. In *IEEE ICDM*, 2010.
[24] Geoffrey H Ball and David J Hall. 1965. Isodata, a novel method of data analysis and pattern classification. Technical report, Stanford Research Inst., Menlo Park, 1965.
[25] N Karthikeyani. 2009. Visalakshi and J Suguna. K-means clustering using max-min distance measure. In *IEEE NAFIPS*, 2009.
[26] Bahman Bahmani et al. 2012. Scalable k-means++. *arXiv:1203.6402*, 2012.
[27] Olivier Bachem et al. 2016. Approximate k-means++ in sublinear time. In *AAAI*, 2016.
[28] Xiaoyan Wang and Yanping Bai. 2016. The global minmax k-means algorithm. *Springerplus*, 5(1):1–15, 2016.
[29] Harlfoxem. 2016. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction, 2016.
[30] Merishna Singh Suwal. 2018. https://www.kaggle.com/datasets/merishnasuwal/breast-cancer-prediction-dataset, 2018.
[31] M Yasser H. 2022. https://www.kaggle.com/datasets/yasserh/wine-quality-dataset, 2022.
[32] Rachael Hwee Ling Sim et al. 2020. Collaborative machine learning with incentive-aware model rewards. In *ICML*, 2020.
[33] Lingjiao Chen et al. 2019. Towards model-based pricing for machine learning in a data marketplace. In *SIGMOD*, 2019.
[34] Alvin E Roth. 1988. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.
[35] Jinsung Yoon et al. 2020. Data valuation using reinforcement learning. In *ICML*, 2020.
[36] Anish Agarwal et al. 2019. A marketplace for data: An algorithmic solution. In *EC*, 2019.
[37] Dongge Han et al. 2020. Replication-robust payoff-allocation for machine learning data markets. *arXiv:2006.14583*, 2020.
[38] Arthur Zimek and Peter Filzmose. 2018. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 8(6):e1280, 2018.
[39] Jianzhou Wang et al. 2020. Outlier-robust hybrid electricity price forecasting model for electricity market management. *J. Clean. Prod.*, 249:119318, 2020.