Natural Selection: A Natural Visibility Approach for Task-Agnostic Time Series Dataset Valuation

Ali Falahati
ali80af@gmail.com

Abstract

Valuing time series datasets effectively is crucial for enhancing machine learning models by selecting high-quality data that improves model performance. In this paper, we propose a task-agnostic method for time series dataset valuation based on a combination of local and global diversity measures. We first represent time series as natural visibility graphs, transforming the data into a graph-based format. We then compute local diversity through Weisfeiler-Lehman graph embeddings and global diversity using the Gaussian Radial Basis Function (RBF) kernel, capturing both local and global structure in the time series data. Our proposed true diversity score, a convex combination of local and global diversity, provides a metric to quantify the intrinsic variability of time series datasets. Additionally, we introduce a relevance metric based on spectral graph entropy and Jensen-Shannon divergence, which measures the similarity between time series datasets. This work contributes to the growing field of data-centric AI by providing tools to assess the value of time series data in a task-agnostic manner, facilitating better decision-making in data selection for machine learning tasks.

1 Introduction

The rapid usage of time series data across domains such as finance, healthcare, climate science, and industrial monitoring has increased the need for machine learning models that can effectively learn from and interpret temporal patterns. However, not all time series datasets contribute equally to model performance. As machine learning transitions from a model-centric paradigm — where performance improvements primarily come from better algorithms — to a *data-centric* approach, the importance of *data valuation* has become paramount. Data-centric AI focuses on improving model outcomes by carefully curating and selecting high-quality datasets, making the ability to quantify the value of data a critical challenge.

Data valuation is the process of determining the worth of a dataset based on its ability to improve model performance, either by enhancing generalization or providing complementary information. While task-specific data valuation approaches have received significant attention, there is growing interest in *task-agnostic* methods — those that can evaluate datasets without depending on the final task or model. This task-agnostic approach is particularly important for time series data, which often exhibit complex temporal dependencies, seasonality, trends, and noise. Existing valuation methods for static datasets often fail to capture these nuances, making them insufficient for time series.

In this paper, we propose a novel task-agnostic framework for *time series dataset valuation*, focusing on two key factors: *diversity* and *relevance*. These complementary aspects offer a comprehensive way to evaluate the quality and utility of datasets, independent of any specific task.

1.1 Diversity

Diversity refers to the range of distinct patterns, behaviors, and structures present within a dataset. A diverse dataset ensures that a machine learning model is exposed to a variety of scenarios, which helps improve generalization, robustness, and adaptability to unseen data. For time series, diversity includes temporal patterns such as trends, seasonality, cyclic behaviors, and fluctuations over time. Diverse datasets enable models to learn more effectively, covering a wider spectrum of possible real-world behaviors.

To quantify diversity in time series datasets, we represent the time series as *natural visibility graphs (NVG)*, converting the temporal data into graph structures. We introduce two types of diversity measures:

- **Local Diversity**: This measure captures fine-grained, short-term variations in the time series by evaluating the differences between small windows of data using graph embeddings. By applying the Weisfeiler-Lehman graph kernel, we assess the differences in the structure of local graphs, which correspond to short-term fluctuations in the time series.
- **Global Diversity**: This measure captures long-term, overarching trends in the time series by comparing the entire dataset. Using the Gaussian Radial Basis Function (RBF) kernel, global diversity captures broad patterns, seasonality, and long-range dependencies.

These two measures are combined to form a *True Diversity* score, which balances short-term variability and long-term structural patterns. This holistic view of diversity ensures that a dataset contains sufficient variety to enrich a machine learning model's training process, ultimately leading to better performance across different tasks.

1.2 Relevance

While diversity captures the richness of a dataset, *relevance* measures how well a dataset aligns with other datasets or complements existing data. In the context of time series valuation, relevance is critical for understanding whether a new dataset offers additional value when combined with already available data. Relevance is particularly important in scenarios such as data marketplaces or iterative model development, where users seek to augment their existing data with complementary datasets.

Our proposed relevance metric is based on the *spectral properties of graph representations* of time series data. By analyzing the spectral entropy of the graphs generated using the NVG algorithm, we capture the complexity of the dataset. We then use *Jensen-Shannon divergence* to compare the spectral densities of two datasets, quantifying the structural differences between them. This approach provides a task-agnostic metric for determining whether new data will add value to a model's training set, independent of the specific task.

Relevance helps answer two key questions in data valuation:

- **Complementarity**: Does a new dataset introduce novel information that complements existing data, potentially improving model performance?
- **Redundancy**: Is the new dataset too similar to the existing one, offering little additional value and increasing the risk of overfitting?

By combining these perspectives on diversity and relevance, our framework provides a robust mechanism for task-agnostic time series dataset valuation. *Diversity* ensures that datasets are varied and comprehensive, while *relevance* helps identify whether new data offers meaningful contributions when added to existing datasets.

In this paper, we make the following key contributions:

- We propose a novel task-agnostic framework for time series dataset valuation based on *natural visibility graphs (NVG)*, offering a graph-based representation that preserves both local and global temporal structures.
- We introduce *True Diversity*, a combination of local and global diversity measures, which quantifies the variability in time series datasets by capturing both short-term and long-term patterns.

- We define a *Relevance* metric grounded in graph spectral entropy and Jensen-Shannon divergence, enabling task-agnostic assessment of how well datasets complement each other.
- We validate our approach through experiments on synthetic and real-world time series datasets, demonstrating that our diversity and relevance metrics effectively guide data selection, resulting in improved model generalization and performance.

This work advances the field of *data-centric AI* by providing a practical, task-agnostic methodology for assessing the value of time series datasets. Our framework facilitates better decision-making in data selection, paving the way for more efficient and effective use of temporal data in machine learning applications.

2 Preliminary

Graph representation: Let a graph G be defined as $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{v_1, \ldots, v_{N_\mathcal{V}}\}$ is the set of nodes, with cardinality $|\mathcal{V}| = N_\mathcal{V}$. $\mathcal{N}(v_i)$ denotes the set of neighboring nodes of v_i, and $w(v_i, v_j)$ represents the weight of the edge between nodes v_i and v_j. G can be represented by an adjacency matrix $A \in \{0,1\}^{N_\mathcal{V} \times N_\mathcal{V}}$, with $A_{ij} = 1$ if nodes v_i and v_j are connected and $A_{ij} = 0$ otherwise.

Random graph: Random graph can be defined as a probability space $(\Omega, \mathcal{F}, P)$, where the sample space Ω is a nonempty set of graphs. The set of events $\mathcal{F}$ is a collection of subsets of Ω, encompassing the power set of Ω, thus including every possible combination of these graphs. The probability measure P assigns a probability to each event in $\mathcal{F}$, quantifying the likelihood of each subset of graphs occurring.

Spectral properties of random graph: Given a set of $N_\mathcal{V}$ labeled nodes $\mathcal{V} = \{v_1, \ldots, v_{N_\mathcal{V}}\}$, let g be a random graph such that its sample space Ω consists of graphs with labnodes $\{v_1, \ldots, v_{N_\mathcal{V}}\}$. We define the *spectrum* of g as a random vector containing $N_\mathcal{V}$ random variables $\lambda_1, \lambda_2, \ldots, \lambda_{N_\mathcal{V}}$. Each function $\lambda_i : \Omega \to \mathbb{R}$ maps a graph in the sample space Ω to the i-th largest eigenvalue of its adjacency matrix.

Let δ be the Dirac delta function, which is the probability measure satisfying

- $\delta(x) = 0, x \in \mathbb{R} \setminus \{0\}$,
- $\delta(0) = \infty$,
- $\int_{-\infty}^{+\infty} \delta(x)\, dx = 1$.

Let G be a graph in the sample space of g. The *empirical spectral density* according to the probability law of g [5] is given by:

$$\rho(\lambda) = \lim_{n_\mathcal{V} \to \infty} \left\langle \frac{1}{N_\mathcal{V}} \sum_{i=1}^{N_\mathcal{V}} \delta\left(\frac{\lambda - \lambda_i}{\sqrt{N_\mathcal{V}}}\right) \right\rangle.$$

3 Graphical Representation of Time Series

Let $T = \{t_1, t_2, \ldots, t_N\}$ represent a time series consisting of N data points, where t_i denotes the data value at time index i. We partition T into N overlapping windows, each of size M, forming a new set of windows T_g as follows:

$$T_g = \{(t_1, t_2, \ldots, t_{M \bmod N}), (t_2, t_3, \ldots, t_{M+1 \bmod N}), \ldots, (t_N, t_1, \ldots, t_{M-1 \bmod N})\}$$

Here, each window contains M consecutive data points from the original series T, with modular arithmetic ensuring that the windows wrap around at the boundary of the time series, making T_g a cyclic or periodic transformation of the original series.

The natural visibility algorithm [3] creates N graphs based on each set $T_{g_i} \in T_g$ and assigns each data point $t_i \in T_{g_k}$ of the window k to a node in the natural visibility graph (NVg). Two nodes i and j in the graph are connected with weight $w(i,j) = |t_i - t_j|$ if one can draw a straight line in the

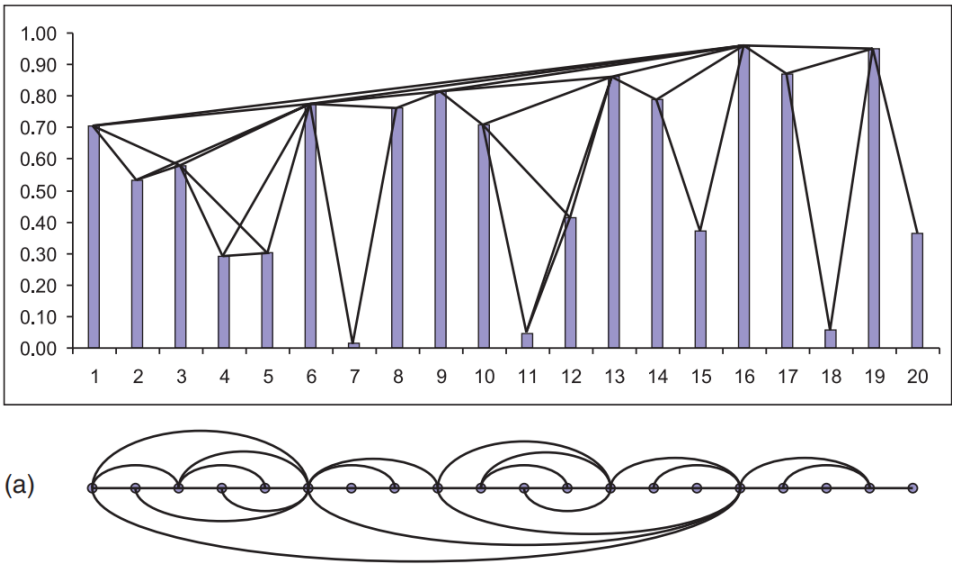

Figure 1: Natural Visibility algorithm

window joining t_i and t_j that does not intersect any intermediate data height t_k. Hence, i and j are two connected nodes if the following geometrical criterion is fulfilled within the window:

$$t_k < t_i + (t_j - t_i)\frac{j-i}{k-i}. \tag{1}$$

Therefore, the result of the NVg would be a set of graphs $\mathcal{G} = \{G_i = (\mathcal{V}_i, \mathcal{E}_i)\}_{i=1,\ldots,N}$ such that $|\mathcal{V}_i| = M$. The graphs extracted with NVg has the following characteristics:

1. **Connected**: Each node is capable of observing at least its nearest neighbors on both the left and right sides.
2. **Undirected**: The algorithm is constructed such that the links have no inherent directionality.
3. **Invariant under affine transformations**: The visibility criterion remains unchanged under (unsigned) linear rescaling of both the horizontal and vertical axes, as well as under horizontal and vertical translations of the series data.

The NVg algorithm is inherently "lossy," meaning that some information from the time series is inevitably lost during the transformation to graphs. We have identified the following key characteristics that are essential for evaluating time series datasets, along with those that are lost during the NVg transformation: (i) Trend, (ii) Seasonality, (iii) Cyclic Patterns, (iv) Stationarity, (v) Noise, and (vi) Volatility. As demonstrated in Table **??** and Appendix **??**, NVg performs well in preserving the information related to the first five characteristics. However, it may suffer from information loss concerning volatility.

Remark: The choice of M significantly affects the amount of information lost in the NVg. Additionally, large values of M can make the computation costly. Therefore, an appropriate value for M should be determined based on the experimental results.

4 Diversity

Diversity in data valuation is crucial for enhancing the performance and reliability of machine learning models. Diverse datasets ensure that models are trained on a wide range of scenarios, leading to better generalization and robustness. In the context of machine learning, diversity helps in capturing the complete structure of the underlying data distribution. Diversity is vital in time series data valuation for several reasons, as it enhances the robustness, accuracy, and generalization capability of models used for forecasting and analysis.

In order to capture the diversity of a time series data we use an approach inspired by Vendi score [1]. The Vendi Score is based on the exponential of the Shannon entropy of the eigenvalues of a similarity matrix, enabling it to measure the effective number of unique elements in a sample. The Vendi Score, though effective for evaluating diversity in static datasets, is not well-suited for time series data due to several key reasons. Time series data inherently involves temporal dependencies where the order and timing of data points are critical, which the Vendi Score's similarity measures do not account for. Additionally, time series data often exhibit trends, seasonality, and non-stationary behavior, which contradict the Vendi Score's assumption of data points being independently and identically distributed. Lastly, the dynamic and evolving nature of time series data, where patterns can change over time, is not captured by the Vendi Score's static similarity matrix.

In order to compromise for these challenges, we introduce two diversity scores. Local diversity and Global diversity.

4.1 Local Diversity

Let $\{G_1, \ldots, G_N\}$ be the graphs extracted from the time series dataset with NVg. For each graph G_i, we have a set of nodes $\{v_1, \ldots, v_M\}$. Let $x(v_i) \in \mathbb{R}$ denote the node attribute. To embed the nodes, we utilize the Weisfeiler-Lehman (WL) scheme. The Weisfeiler-Lehman subtree kernel [6, 7] examines similarities among subtree patterns through a propagation scheme that iteratively updates node attributes based on their neighbors. We define the initial attribute $x^0(v_i) = x(v_i)$ for each node v_i. Let H be the number of WL iterations. For every $h \in \{0, \ldots, H\}$ recursively, we define

$$x^{h+1}(v_i) = \frac{1}{2}\left(x^h(v_i) + \frac{1}{\deg(v_i)} \sum_{u \in \mathcal{N}(v_i)} w((v_i, u)) \cdot x^h(u)\right).$$

As the updating process for the WL. The WL features are defined as

$$X_G^h = \begin{bmatrix} x^h(v_1) \\ \vdots \\ x^h(v_{n_G}) \end{bmatrix},$$

where X_G^h is a column vector of node attributes at iteration h. The final node embeddings of graph G at iteration H are defined as

$$X_G^H := \text{concatenate}(X_G^0, \ldots, X_G^{H-1}).$$

Based on [8], we define the Graph Wasserstein Distance (GWD) as follows:

Definition (Graph Wasserstein Distance). Given two graphs $G = (\mathcal{V}, \mathcal{E})$ and $G' = (\mathcal{V}', \mathcal{E}')$ with respective node embeddings at iteration H, $X = X_G^H$ and $X' = X_{G'}^H$, we define the Graph wasserstein distance (GWD) as

$$W(G, G') := \min_{P \in \Gamma(X, X')} \langle P, D \rangle.$$

Here, D is the distance matrix containing the distances $d(x, x') = \|x - x'\|_2$ between each element x of X and x' of X', $P \in \Gamma$ is a transport matrix, and $\langle \cdot, \cdot \rangle$ is the Frobenius dot product. The transport matrix P contains the fractions that indicate how to transport the values from X to X' with the minimal total transport effort.

We now define the local diversity kernel k_l for a pair of graphs $G = (\mathcal{V}, \mathcal{E})$ and $G' = (\mathcal{V}', \mathcal{E}')$ as

$$k^l(G, G') = e^{-\beta W(G, G')}.$$

Where β is a hyperparameter. Given the graph set $\mathcal{G} = \{G_1, \ldots, G_N\}$ extracted from the time series dataset T, we compute the local similarity matrix $S^l \in R^{N \times N}$ as

$$S^l_{i,j} = k^l(G_i, G_j).$$

4.2 Global Diversity

While the local diversity measure can capture the diversity and patterns within the graph sets—effectively dealing with the intricacies of the graphs themselves—the global diversity measure calculates the diversity based solely on the inherent characteristics of the time series data points. This global approach provides a comprehensive diversity measure that is robust to local variations and noise within individual time series. By focusing on the overall structure and distribution of the data points across the entire dataset, the global diversity measure can identify broad patterns and trends that may be overlooked by local measures.

To compute the similarity between time series points $T = \{t(1), ..., t(N)\}$, we utilize the Gaussian Radial Basis Function (RBF) kernel. The Gaussian RBF kernel examines the similarities among the points through a Gaussian function. We define the kernel function as

$$k^g(t(i), t(j)) = \exp\left(-\frac{\|t(i) - t(j)\|^2}{2\sigma^2}\right),$$

where σ is the bandwidth parameter that controls the width of the Gaussian function.

Given the time series $T = \{t(1), ..., t(N)\}$, we compute the global similarity matrix $S^g \in R^{N\times N}$ as

$$S^g_{i,j} = k^g(t(i), t(j)).$$

4.3 True Diversity

So far, we have calculated the local and global similarity matrices but how can we use them in order to calculate the diversity of a time series dataset? Inspired by ecology, diversity is often defined as the exponential entropy of a species distribution, a concept that captures the variety within a population and decreases as the distribution becomes less uniform [2, 4]. Vendi score [1] extended the idea to machine learning by solely considering samples and defining diversity as the exponential of the Shannon entropy of a similarity matrix over the samples. Therefore we define true diversity as follows:

Definition (True Diversity). Let $\{\lambda^l_1, \ldots, \lambda^l_N\}$ be the eigenvalues of $S^l/N \in \mathbb{R}^{N\times N}$ and $\{\lambda^g_1, \ldots, \lambda^g_N\}$ be the eigenvalues of $S^g/N \in \mathbb{R}^{N\times N}$. True Diversity ($\mathcal{D}$) for time series T is defined as the convex combination of the exponential of the Shannon entropy of both eigenvalues from global and local similarity matrix

$$\mathcal{D}(T) = \alpha \exp\left(-\sum_{i=1}^{n} \lambda^l_i \log \lambda^l_i\right) + (1-\alpha)\exp\left(-\sum_{i=1}^{n} \lambda^g_i \log \lambda^g_i\right).$$

Where α is a hyperparameter.

5 Relevance

To determine the relevance between two time series, we must first define what it means for two time series to be relevant. Previously, we assessed diversity using graphical representations of the time series data. Thus, it is necessary to find a solution within the space of graphs to unify the concepts of diversity and relevance. We define relevance as the amount of structural difference between two graph sets extracted with NVg from time series datasets. It is important to note that we are dealing with sets of graphs rather than individual graphs. To proceed, we treat each graph $G \in \mathcal{G}$ as generated by a random graph model. Therefore, we are interested in quantifying the difference between two random graphs.

The entropy of a random graph provides a measure of the randomness in its structure, making it suitable for this purpose. Entropy captures the complexity and variability within the graph's structure, reflecting how predictable or unpredictable the graph is. By comparing the entropies of two random graphs, we can assess the differences in their structural complexities. This approach leverages the probabilistic nature of random graphs and the comprehensive structural information encapsulated in

the graph's spectral properties. Let g be a random graph with spectral density ρ. The spectral entropy of g is defined as

$$\mathcal{H}(\rho) = -\int_{-\infty}^{+\infty} \rho(\lambda) \log \rho(\lambda)\, d\lambda.$$

Calculating $\mathcal{H}(\rho)$ explicitly is often infeasible, so we will approximate it. For a given graph set $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ with N nodes, for each G_j, $1 \leq j \leq N$, we apply a density function estimator based on the Gaussian kernel in order to estimate the spectral density. Given a graph G_j and its spectrum $\{\lambda_1^{(j)}, \lambda_2^{(j)}, \ldots, \lambda_n^{(j)}\}$, each eigenvalue λ_i contributes to estimate the function in a point λ according to the difference between λ_i and λ. That contribution is weighted by the kernel (K) function and depends on a parameter known as bandwidth (h), which controls the size of the neighborhood around λ. Formally, the density function estimator at a point λ is

$$\hat{f}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{\lambda - \lambda_i}{h}\right),$$

where

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

To obtain an estimator for the random graph, we apply the procedure described above for each observed graph $\{G_1, G_2, \ldots, G_N\}$ and then take the average among all the estimators.

Now that we have estimated the spectral entropy of the random graphs, we return to our main question: how do we compare two graph sets $\mathcal{G}_1$ and $\mathcal{G}_2$? We begin by treating each graph set as a random graph, denoted g_1 and g_2, respectively. Next, we estimate their spectral densities ρ_1 and ρ_2 using the previously mentioned procedure. Before defining relevance, we have to define the Kullback–Leibler (KL) divergence for the random graphs. Let g_1 and g_2 be two random graphs with spectral densities ρ_1 and ρ_2, respectively. The KL divergence is defined as follows. If the support of ρ_2 contains the support of ρ_1, then the KL divergence between ρ_1 and ρ_2 is

$$KL(\rho_1 \| \rho_2) = \int_{-\infty}^{+\infty} \rho_1(\lambda) \log \frac{\rho_1(\lambda)}{\rho_2(\lambda)}\, d\lambda,$$

We define relevance as follows:

Definition (Relevance). Let ρ_1 and ρ_2 be the spectral densities estimated from the graph sets $\mathcal{G}_1$ and $\mathcal{G}_2$ which derived from time series datasets T_1 and T_2. We define relevance as the Jensen–Shannon divergence between two spectral densities

$$\mathcal{R}(T_1, T_2) = \frac{1}{2} KL(\rho_1 \| \rho_M) + \frac{1}{2} KL(\rho_2 \| \rho_M),$$

where $\rho_M = \frac{1}{2}(\rho_1 + \rho_2)$.

The Jensen–Shannon divergence can be interpreted as a measure of the structural differences between two random graphs, quantifying their relevance.

6 Experiments

Our experimental evaluation focuses on three key aspects of our work: (i) We demonstrate that our introduced metric relevance effectively discriminates between time series datasets from different contexts, capturing the essence of relevance. This evaluation validates the metric's ability to quantify the pertinence of datasets to specific tasks or domains. (ii) We assess the efficacy of our proposed diversity metric across various synthetic and real-world settings. This analysis showcases how the metric captures and quantifies the diversity within and between datasets, providing insights into dataset composition and potential information gain. (iii) We test our approach in practical scenarios involving multiple datasets offered to a buyer who has already trained a model with an existing dataset. This experiment simulates a marketplace where the buyer seeks the optimal complementary dataset from various sellers. We evaluate how well our metrics guide the selection of datasets that best augment the buyer's existing data, leading to improved model performance.

References

[1] Dan Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*, 2023.

[2] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[3] Lucas Lacasa, Bartolo Luque, Fernando Ballesteros, Jordi Luque, and Juan Carlos Nuno. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972–4975, 2008.

[4] Tom Leinster. Entropy and diversity: The axiomatic approach. *arXiv preprint arXiv:2012.02113*, 2020.

[5] Timothy Rogers. *New results on the spectral density of random matrices*. PhD thesis, King's College London, 2010.

[6] Nino Shervashidze and Karsten Borgwardt. Fast subtree kernels on graphs. *Advances in neural information processing systems*, 22, 2009.

[7] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

[8] Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. Wasserstein weisfeiler-lehman graph kernels. *Advances in neural information processing systems*, 32, 2019.