Evaluating the Reliability of Shapley Value Estimates: An Interval-Based Approach

Evaluating the Reliability of Shapley Value Estimates: An Interval-Based Approach

Davide Napolitano[1], Luca Cagliero[1]

[1]Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

Abstract

Shapley Values (SVs) are concepts used in game theory that have recently found application in Artificial Intelligence. They are exploited to explain models by quantifying the separate features' contribution to the predictor estimates. However, the reliability of the estimated SVs is often not thoroughly assessed. In this context, we leverage Interval Shapley Values (ISVs) to evaluate the importance and reliability of features' contributions when the classifier consists of an ensemble method. This paper presents a suite of ISVs estimators based on exact estimation, linear regression, and Monte Carlo sampling. In detail, we adapt classical SVs estimators to ISV-like concepts to efficiently handle real tabular datasets. We also provide a set of ad hoc performance metrics and visualization techniques that can be used to explore models' results under multiple aspects.

Keywords

Explainable Artificial Intelligence, Interval Shapley Values, Feature Importance

1. Introduction

Shapley Values (SVs), originally formulated in coalition game theory [1], are now widely used to generate post-hoc explanations for classifiers that assign discrete classes to unlabeled samples. In detail, SVs quantify the contribution of each input feature to a given classifier's prediction and, although they may not always accurately reflect feature importance [2], these contributions can be estimated on a per-sample basis (locally) or aggregated to provide insights into the overall behavior of the model (globally) [3].

When a model comprises multiple predictors, estimating the contributions of individual features becomes challenging, as each feature may influence each predictor differently. In some cases, certain predictors might entirely disregard features crucial to others. This implies that the performance provided by the various predictors can vary substantially, directly reflecting on the contributions made by the various features. Therefore, taking into account the contribution of each predictor makes the explanations robust to variability in the estimates (see Figure 2).

To model the variability of SVs across multiple predictors, we rely on the concept of Interval Shapley Values (ISVs) [4]. Derived from the field of cooperative interval games, they can be used to estimate SVs in the presence of uncertainty by encompassing different predictor outcomes, which are neglected in standard SVs. To ensure tractable and scalable computation on real data, we focus on Interval Shapley-Like Values (ISLVs), known to approximate ISVs [5, 6].

HI-AI@KDD, Human-Interpretable AI Workshop at the KDD 2024, 26th of August 2024, Barcelona, Spain

davide.napolitano@polito.it (D. Napolitano); luca.cagliero@polito.it (L. Cagliero)

Hereafter, we present a suite of algorithms adapted to explain combinations of predictors with ISLVs. They indicate the features' importance for the ensemble method's outcomes by explicitly indicating the reliability of such estimates. This is crucial to trust models' explanations and compare the outcomes of different estimators. The suite includes approaches to SVs estimation adapted to handle Interval-based scenarios successfully. Specifically, differently from the neural approaches proposed in [7], we focus on a linear regressor, a Monte Carlo sampling strategy, and an Exact estimator, aiming to incorporate all implementations in BONES [8] library. To allow end-users to explore and compare the outcomes of Interval-based approaches, the suite supports ad hoc performance metrics, extended from the standard SVs scenario to support interval-level evaluations. The metrics can be visualized to ease model comparisons and complexity analysis.

The remainder of this paper is organized as follows. Section 2 introduces the preliminary notions. Section 3 describes the suite of Interval-based approaches. Section 4 shows examples of outcomes and comparisons. Finally, Section 5 draws the conclusions of the work.

2. Preliminaries

In a cooperative game, the Shapley Value ϕ_i represents the contribution of a single player i to the total payoff of a group of player P [9], where ϕ_i is equal to the sum of the weighted marginal contributions of i to P over all possible player's coalitions $S \subseteq P$. Beyond explaining an individual sample x, Shapley Values can be leveraged to provide a global explanation of the dataset by averaging sample-level contributions [10, 11].

Suppose to have the outcome pr_x of an ensemble $\mathbf{M}$ of predictors on a sample x with a confidence interval $[\underline{pr}_x, \overline{pr}_x]$. In compliance with [4], we define the *Coalitional Interval Game* [12, 13] as a pair (w, S), where w: $2^P \rightarrow I(\mathbb{R})$is a function that maps an arbitrary coalition $S \subseteq P$ to the corresponding confidence interval $w(S)$: $[\underline{w}(S), \overline{w}(S)] = [\underline{pr}(S), \overline{pr}(S)]$.

To explain ensemble methods we use the concept of Interval Shapley Values (ISVs) [4] associated with each Coalitional Interval Game (w, P) to a payoff vector where each component is a compact interval of real numbers [14]. In a nutshell, ISVs capture the range of contributions of a feature p by evaluating the interval values across all possible feature combinations. ISVs have to satisfy two notable properties:

- **Partial Subtractor:** given two intervals I and J, the Partial Subtraction $I - J$ is defined as $[\underline{I} - \underline{J}, \overline{I} - \overline{J}]$ only if $\Delta_I \geq \Delta_J$, where Δ is the interval width: $I = [\underline{I}, \overline{I}] \rightarrow \Delta = \overline{I} - \underline{I}$.
- **Size Monotonicity:** ISVs can be defined only when the Coalitional Interval Game (w, P) is *size monotonic*, i.e., when $\Delta w(S) \leq \Delta w(T)$ for all $S, T \in 2^P$ with $S \subset T$.

Since the ISVs constraints are computationally intractable [9, 15], Interval Shapley-Like Values [15] offer a more efficient yet approximated approach to ISVs estimation. ISLVs adopt the Moore operators [16], in detail the Moore Subrtractor is used rather than the Partial Subtractor operator, i.e., given two intervals I and J the Moore subtraction is defined as $I \ominus J = [\underline{I} - \overline{J}, \overline{I} - \underline{J}]$. To simplify the estimation [15], the *Median* and *Uncertain-Spread* games are introduced:

- *Median Game* (w_m, P): $$w_m(S) = [\frac{\underline{w(S)} + \underline{\overline{w(S)}}}{2}, \frac{\overline{w(S)} + \overline{\overline{w(S)}}}{2}], S \in 2^P \quad (1)$$
- *Uncertain-Spread Game* (w_u, P): $$w_u(S) = [\frac{-\Delta w(S)}{2}, \frac{\Delta w(S)}{2}], S \in 2^P \quad (2)$$

Hereafter, we focus on two ISLVs definitions based on the *Median* and *Uncertain-Spread* games:

- **Improved ISLVs [15]**: $$I\Phi_i^I(w) = \Phi_p^I(w_m) \oplus \frac{\Delta\Phi_i^I(w_u)}{\sum_{p\in P}\Delta\Phi_i^I(w_u)} w_u(P) \quad (3)$$
- **Reformulated ISLVs [6]**: $$R\Phi_i^I(w) = \Phi_i^I(w_m) \oplus \frac{1}{|P|} w_u(P) \quad (4)$$

where $\oplus$ is the Moore Addition $I \oplus J = [\underline{I} - \underline{J}, \overline{I} - \overline{J}]$.

3. Suite of Interval-based Approaches

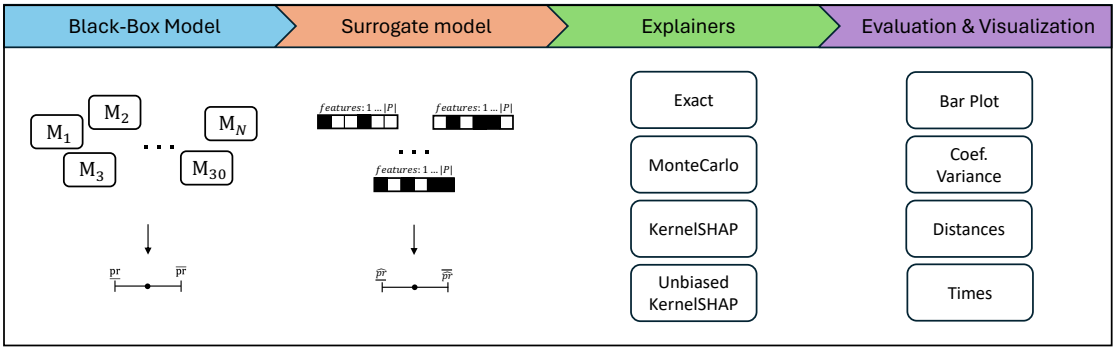

Figure 1: Schema of the presented suite

We present a suite of SVs estimators adapted to handle Interval-based estimations on tabular data. To successfully explain ensembles of predictors, the suite integrates adaptations of existing algorithms that produce Improved [15] and Reformulated [6] estimates of ISLVs instead of classical SVs. The suite is available for research at the link: https://github.com/kddwshi/KddHI.

A sketch of the suite is depicted in Figure 1.The black-box model $\mathbf{M}$ to be explained consists of an ensemble of N independent predictors, which are all trained on a labeled relational dataset. For every instance x to be classified, each predictor returns its corresponding per-class output probabilities, which are then used to compute the confidence interval to retrieve an interval payoff for the ensemble method. As discussed in the Preliminaries, the standard formulations of both SVs and ISVs involve evaluating model contributions across different subsets of features. Since most of the existing models do not support holding out subsets of features, similar to [7, 17], we exploit a surrogate model to approximate the original model considering subsets of features, thus allowing the subsequent normalization of ISLVs similar to [11, 18].

To adapt traditional SVs-based explainers to ISLVs, we leverage the *Median* and *Uncertain-Spread* games according to the Improved [15] and Reformulated [6] ISLVs formulations.

Median game The ISLVs can be expressed as a single value since the characteristic function w_m is defined as an interval with equal interval endpoints. This approach allows the estimation to be performed using established methods. Subsequently, the interval can be reconstructed in the next step when defining the ISLVs.

Uncertain-Spread game Since the minimum and maximum values returned by the characteristic function w_u are opposites (i.e., same value, opposite sign), the ISLVs estimation can be simplified by applying the addition operation, rather than the subtraction, upon the single absolute value. This consideration is exemplified in Equation 5, where the subtraction is reconducted to the addition of the absolute values retrieved from different subsets S applied on w_u.

$$\begin{aligned} &w_u(S_1, x) = [-v_1, v_1], w_u(S_2, x) = [-v_2, v_2] \\ &w_u(S_1, x) \ominus w_u(S_2, x) = [-v_1 - v_2, v_1 + v_2] = [-(v_1 + v_2), v_1 + v_2] \end{aligned} \tag{5}$$

Therefore, since managing the computation with a single value, we can reconduct the estimation to the traditional SVs formulation. In this way, classical predictors can be directly exploited to retrieve the absolute values and, following, to reconstruct the desired $\Phi_i^I(w_u)$.

Based on the considerations above, we adapt the following algorithms to support ISLVs: the Exact explainer [11], Unbiased and Biased KernelSHAP [19] and Monte Carlo sampling [20]. For each algorithm, we separately implement adaptations based on *Median* and an *Uncertain-Spread*, namely the Improved [15] and Reformulated [6] versions.

3.1. Performance metrics

Given the algorithms' outcomes achieved on a relational dataset, the suite allows the quantitative evaluation of (1) The accuracy of the intervals estimated by each algorithm against a ground truth in terms of (a) the $L2$ distance between the mean points, or (b) the $L2$ distance between the interval widths, or (c) the Euclidean distance between the intervals [21]. (2) The efficiency of the estimators in terms of training and inference time. Whenever not otherwise specified, we use the Exact algorithm adaptation as reference ground truth.

3.2. Outcome visualizations

The suite supports the following graphical visualization of the experimental results achieved on a test dataset: (1) A bar plot showing the per-feature intervals, which may allow a direct comparison between different algorithms; (2) A graph plotting the coefficient of variation of the ISVs (width over mean point), which provides insights into the reliability of the generated estimated; (3) A plot showing the computational times for model training and inference by varying the dataset size and dimensionality.

4. Preliminary results

We show examples of outcomes achieved on four relational datasets taken from the UCI repository [22] namely Monks, Bank, Wisconsin Breast Cancer, and Diabetes. We explain a Random

Forest Classifier with 100 predictor trees, implemented in the Scikit-learn library [23]. We generate the confidence interval (with confidence level $\gamma = 0.95$) from the prediction of each tree. Then, we approximate the predictions of the black-box model using a Multi-Layer Perceptron (MLP) as a surrogate model. Similar to [17, 7], MLP consists of three linear layers, each one with a hidden size of 512 units, interspersed with Rectified Linear Unit (ReLU) activation functions, and with two final classification heads. The surrogate model was trained for up to 200 epochs using the Kullback-Leibler divergence loss function. The training utilized the AdamW optimizer [24], with a learning rate of 10^{-4}, a batch size of 8, and a weight decay of 10^{-2}.

Regarding the explainers implementation, the baselines of *Median* and *Uncertain-Spread* Exact explainers are trained on 100 samples. Concerning the Monte Carlo approach, the number of iterations is set to 1000. For the KernelSHAP-based methodologies, we adopt the marginal models' approach as outlined in [25]. Specifically, the *Median* marginal model was configured with 20 baseline samples, while the *Uncertain-Spread* marginal model was allocated 8 baseline samples. These sample sizes were carefully chosen to strike a balance between achieving accurate estimations and maintaining computational efficiency. Indeed, higher values lead to comparable results but with longer times, while lower values, although providing shorter times, give worse estimates. Moreover, regarding the iteration parameters of the two regression-based methods, the results are retrieved by testing all datasets with a threshold of 0.1 and a kernel iteration value of 128.

4.1. Examples of results and visualizations

Table 1
ISLVs Estimators - $L2$ Distances on Interval Mean values ($L2_M$) and on Interval Withds value ($L2_W$).

	$L2_M$	$L2_W$	$L2_M$	$L2_W$	$L2_M$	$L2_W$	$L2_M$	$L2_W$
	Bank		**Monks**		**Diabetes**		**WBC**	
I-UKS	**0.049±0.001**	0.003±0.001	0.026±0.001	**0.019±0.001**	0.024±0.001	**0.019±0.002**	**0.028±0.001**	**0.008±0.001**
I-KS	0.050±0.002	0.003±0.001	0.026±0.001	0.021±0.003	0.025±0.002	0.026±0.002	0.034±0.003	0.011±0.001
I-MC	0.076±0.004	**0.001±0.001**	**0.016±0.001**	0.036±0.001	**0.021±0.001**	0.022±0.001	0.034±0.002	**0.008±0.001**

In this section, we present a comparative analysis of the proposed models based on various metrics. In detail, the table results are shown as confidence intervals computed on 5 different runs with a machine equipped with an AMD Ryzen 7950X CPU. Table 1 illustrates the comparison of ISLVs with respect to the mean point and interval width. The results indicate that the outputs of Unbiased KernelSHAP and Monte Carlo Sampling best approximate the Exact model, with Unbiased KernelSHAP yielding superior results in terms of amplitude precision.

Moreover, Table 1 reports the results exclusively for the Improved models (denoted with the prefix *I-*), as they share mean points with the Reformulated models (denoted with the prefix *R-*), and the interval amplitudes for the latter remain invariant regardless of the approach. Similar takeaways can be derived from examining the Euclidean distances of the intervals presented in Table 2, where the Improved and Reformulated approaches yield similar rankings.

Table 2
ISLVs Estimators - Euclidean distances between intervals

	R-UKS	R-KS	R-MC	I-UKS	I-KS	I-MC	R-Exact vs I-Exact
Bank	**0.0698±0.0005**	0.0710±0.0031	0.1073±0.0052	**0.0702±0.0005**	0.0713±0.0031	0.1074±0.0052	0.0003±0.0001
Monks	0.0371±0.0001	0.0372±0.0003	**0.0291±0.0019**	**0.0393±0.0003**	0.0481±0.0025	0.0523±0.0002	0.0001±0.0001
Diabetes	0.0337±0.0008	0.0349±0.0030	**0.0292±0.0014**	0.0449±0.0022	0.0530±0.0005	**0.0446±0.0008**	0.0005±0.0001
WBC	**0.0396±0.0009**	0.0481±0.0041	0.0476±0.0023	**0.0434±0.0005**	0.0529±0.0043	0.0508±0.0020	0.0005±0.0001

Figure 2 provides a visual insight into the numerical results, i.e., a bar plots and a Coefficient of Variation plot computed using Unbiased KernelSHAP on the Diabetes dataset. This example demonstrates how the interval data can be employed to assess the reliability of the Shapley Values associated with each feature.

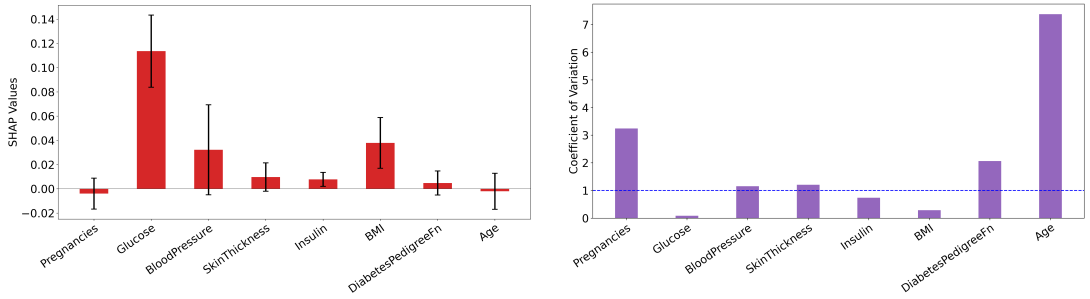

Figure 2: Bar Plot and Coefficient of Variation Plot on the Diabetes dataset. Generally, when the amplitude is substantially larger than the midpoint (e.g., coefficient of variation > 1), the reliability of the feature estimate should be carefully reconsidered.

4.2. Execution times

Table 3 compares the inference times per sample spent by all analyzed approaches separately for each tested dataset. The reported statistics show that the Reformulated ISLVs are overall faster compared to the Improved ones. Moreover, the results show that the Reformulated regression-based approach outperforms the other ones when the number of features increases.

Table 3
Average inference times for each model and dataset.

	N° Features	R-Exact	R-UKS	R-MC	I-Exact	I-UKS	I-MC
Bank	4	**0.727±0.118**	4.104±0.457	3.687±0.102	**1.455±0.218**	10.874±2.106	7.374±0.205
Monks	6	**0.456±0.047**	1.479±0.033	1.818±0.047	**0.913±0.085**	14.982±0.737	3.635±0.094
Diabetes	8	2.960±0.028	**1.936±0.007**	2.443±0.022	5.950±0.049	7.022±1.178	**4.887±0.044**
WBC	9	3.084±0.090	**2.148±0.074**	2.768±0.069	6.202±0.191	11.808±1.407	**5.536±0.137**

Summarizing the results, the Reformulated Unbiased KernelSHAP and Monte Carlo approaches yield comparable outcomes on the distances, with the former being favored for

Improved ISLVs. Furthermore, considering inference times, the Reformulated Unbiased KernelSHAP method provides the best overall results, especially as the number of features in the dataset increases.

5. Conclusions and future developments

The paper presented a suite of SVs estimators adapted to explain ensembles of predictors using ISVs. To estimate both importance and reliability of the features' contributions to the black-box model estimates, we adapt three classical SV estimators to handle Intervals of Shapley Values by leveraging the concepts of Interval Shapley-Like Values. The suite allows researchers and practitioners to interact with Interval-based approaches and evaluate them using ad hoc performance metrics and visualizations.

In future work, we plan to investigate approaches not relying on surrogate models, to analyze new sampling techniques, and, most importantly, to extend this technique to other data modalities and in multimodal analyses, such as text and images combined.

References

[1] L. S. Shapley, A value for n-person games, in: H. W. Kuhn, A. W. Tucker (Eds.), Contributions to the Theory of Games II, Princeton University Press, Princeton, 1953, pp. 307–317.

[2] X. Huang, J. Marques-Silva, The inadequacy of shapley values for explainability, arXiv preprint arXiv:2302.08160 (2023).

[3] W. Saeed, C. Omlin, Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems 263 (2023) 110273.

[4] S. Alparslan Gök, R. Branzei, S. Tijs, The interval shapley value: an axiomatization, Central European Journal of Operations Research 18 (2010) 131–140.

[5] S. Ishihara, J. Shino, Some properties of interval shapley values: An axiomatic analysis, Games 14 (2023) 50.

[6] W. Feng, W. Han, Z. Pan, A reformulated shapley-like value for cooperative games with interval payoffs, Operations Research Letters 48 (2020) 758–762.

[7] D. Napolitano, L. Vaiani, L. Cagliero, Efficient neural network-based estimation of interval shapley values, IEEE Transactions on Knowledge and Data Engineering (2024).

[8] D. Napolitano, L. Cagliero, Bones: a benchmark for neural estimation of shapley values, 2024. URL: https://arxiv.org/abs/2407.16482. arXiv:2407.16482.

[9] L. S. Shapley, Notes on the N-Person Game II: The Value of an N-Person Game, RAND Corporation, Santa Monica, CA, 1951.

[10] C. Frye, C. Rowat, I. Feige, Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability, arXiv preprint arXiv:1910.06358 (2019).

[11] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017, pp. 4765–4774.

[12] S. Alparslan Gök, R. Branzei, S. Tijs, Convex interval games, Journal of Applied Mathematics and Decision Sciences 2009 (2009).
[13] S. Z. Alparslan Gök, Cooperative interval games (2009).
[14] L. Carpente, B. Casas-Méndez, I. García-Jurado, A. van den Nouweland, Coalitional interval games for strategic games in which players cooperate, Theory and Decision (2008) 253–269.
[15] W. Han, H. Sun, G. Xu, A new approach of cooperative interval games: The interval core and shapley value revisited, Operations Research Letters 40 (2012) 462–468.
[16] R. E. Moore, Methods and applications of interval analysis, SIAM, 1979.
[17] D. Napolitano, L. Vaiani, L. Cagliero, et al., Learning confidence intervals for feature importance: A fast shapley-based approach, in: Workshop Proceedings of the EDBT/ICDT 2023 Joint Conference (March 28-March 31, 2023, Ioannina, Greece), 2023.
[18] N. Jethani, M. Sudarshan, I. C. Covert, S. Lee, R. Ranganath, Fastshap: Real-time shapley value estimation, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022.
[19] I. Covert, S.-I. Lee, Improving kernelshap: Practical shapley value estimation via linear regression, arXiv preprint arXiv:2012.01536 (2020).
[20] E. sSrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowledge and Information Systems 41 (2014) 647–665.
[21] O. Kosheleva, V. Kreinovich, Euclidean distance between intervals is the only representation-invariant one (2020).
[22] K. Bache, M. Lichman, UCI machine learning repository, 2013.
[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[24] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
[25] I. Covert, C. Kim, S.-I. Lee, Learning to estimate shapley values with vision transformers, 2023. arXiv:2206.05282.