Shaping Up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection

Gwladys Kelodjou[1], Laurence Rozé[2], Véronique Masson[1], Luis Galárraga[1], Romaric Gaudel[1], Maurice Tchuente[3], Alexandre Termier[1]

[1]Univ Rennes, Inria, CNRS, IRISA - UMR 6074 | [2]Univ Rennes, INSA Rennes, CNRS, Inria, IRISA - UMR 6074 | [3]Sorbonne University, IRD, University of Yaoundé I, UMI 209 UMMISCO

Context

- Need to explain Machine Learning decisions for opaque algorithms.
- Popular local post-hoc agnostic methods: **SHAP** [1], LIME [2], DeepLIFT [3], etc.
- They suffer from **stability issues**.

SHAP (Shapley Additive Explanations)

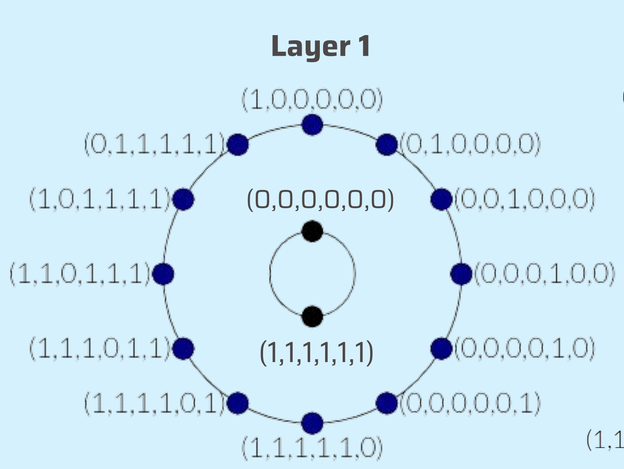
- Applies cooperative game theory to reveal feature contributions.
- **Cooperative game theory**: fairly distribute the total payoff obtained by multiple players collaborating in a game.
- **Shapley value** [4]: Consider different coalitions in which the player is present or absent to determine his contribution.
- SHAP: the players are the features used by the model and the gain is the model's prediction.
- **Kernel SHAP**: model-agnostic approximation of SHAP values based on linear regression.

Problem Statement

How to get stable explanations, verifying properties close to the original SHAP values ?

Kernel SHAP: Background

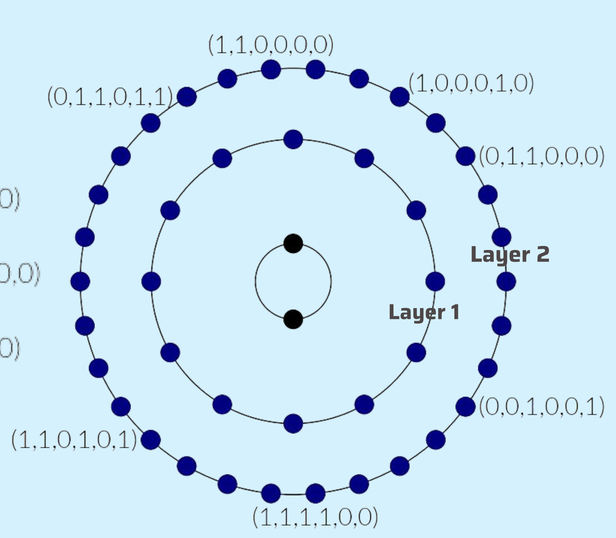
- **Coalition**: subset of features represented by a binary vector, where the presence of a feature is indicated by 1 and its absence is indicated by 0.
- **Layer**: Set of coalitions sharing the same number of features present or features absent.



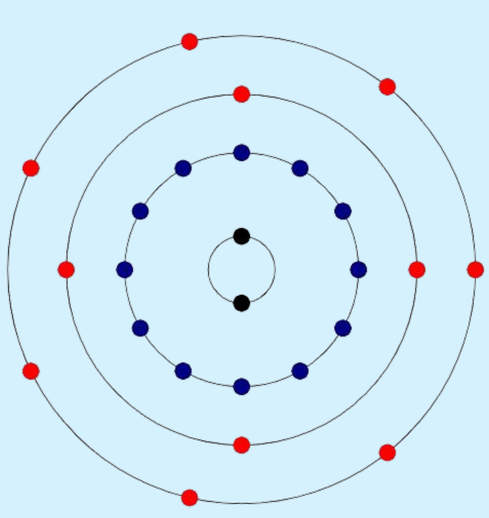
Stratum: combination of successive layers.

Kernel SHAP: Neighbor Generation Process

- Samples a subset of coalitions based on a **budget**.
- First generate coalitions from lower layers **(blue)** and randomly samples from subsequent layers **(red)** if the budget is not exhausted.



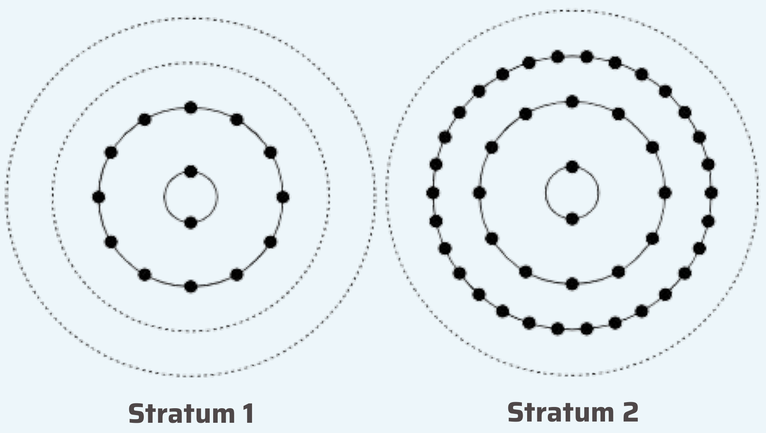
Stratum 2



Random sampling within layers 2 and 3

Achieving Kernel SHAP's Stability

Set the budget to consider only complete stratums.

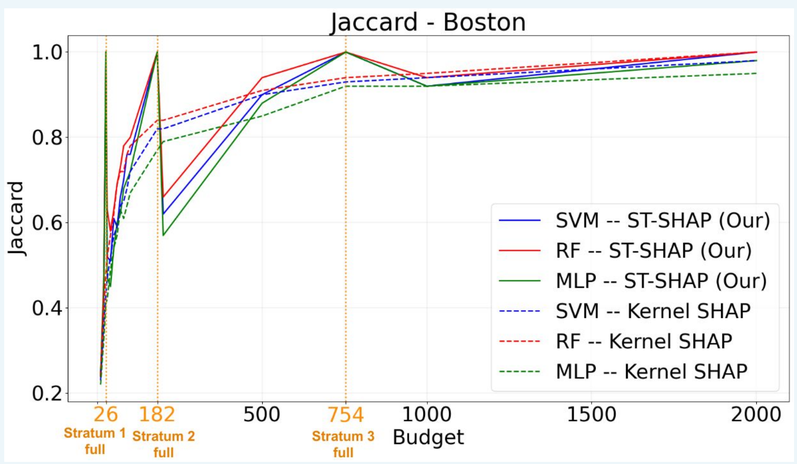


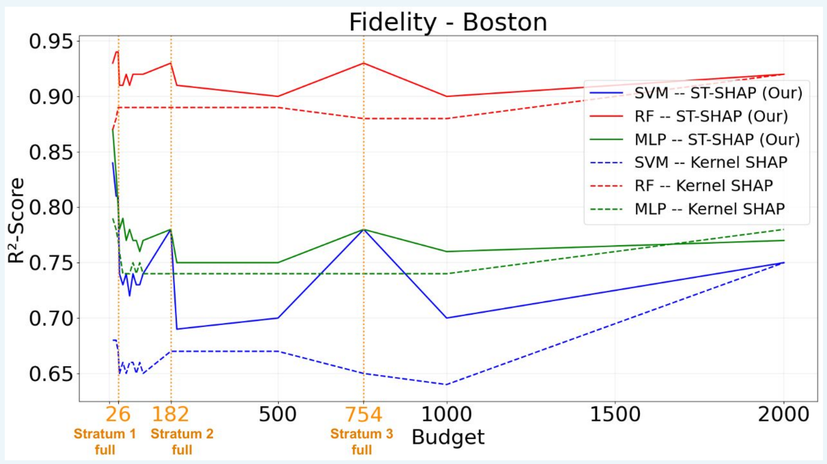
Stratum 1 **Stratum 2**

Complete stratums ⟹ stable output.

Experimental Protocol [5]

- **Black-box models**: SVM, Random Forest (RF), Logistic Regression (LR), Multi-layer Perceptron (MLP).
- **Datasets**: Boston, Movie, Adult, Dry Bean, Default of Credit Card Clients, HELOC, Spambase, Wisconsin Diagnostic Breast Cancer.
  - 10 instances in the test set of each dataset.
- **Various budgets** to assess **stability** and **fidelity**.
- **Metrics**:
  - **Jaccard coefficient** for stability measurement.
    - Repeating the computation of the explanation 20 times for each explained instance.
    - Explanation size: 4 (number of non-zero coefficients returned).
  - **R^2-Score** and **Accuracy** for fidelity measurement.





No impact on explanation fidelity observed with our methods: remains highly accurate.

Stratum 1 Attribution Values

For any feature $j \in N = \{1, \cdots, M\}$ (the set of all features), attribution ϕ_j with Stratum 1 is:

$$\phi_j = \tilde{\phi}_j + \frac{1}{M}\left(f(N) - f(\emptyset) - \sum_{i=1}^{M} \tilde{\phi}_i\right)$$

where for any i, $\tilde{\phi}_i = \frac{f(\{i\}) - f(\emptyset) + f(N) - f(N\backslash\{i\})}{2}$,
with f being the black-box model, and M representing the total number of features.

Stratum 1 Attributions vs SHAP: Properties

Stratum 1	SHAP values

• LES family [6][7]:
  ○ Linearity
  ○ Efficiency
  ○ Symmetry
• Missingness
• **Execution time: O(M)**
	• Linearity
	• Efficiency
	• Symmetry
	• Missingness
	• Null effect
	• **Execution time: $O(2^M)$**

Stratum 1 Attributions vs SHAP: Values

Experiments are conducted on a subset of the datasets for which computing the exact SHAP values is feasible.

		Kendall τ			R^2-Score		
		SVM	RF	MLP	SVM	RF	MLP
Boston	Mean	0.95	0.91	0.959	0.98	0.99	0.99
	Median	0.97	0.92	0.97	0.99	0.99	0.99
Adult	Mean	0.7	0.68	0.87	1.0	0.65	0.41
	Median	0.6	0.7	0.9	1.0	0.77	0.69
Dry Bean	Mean	0.86	0.75	0.84	−0.07	0.74	0.75
	Median	0.9	0.76	0.88	0.51	0.79	0.8
Movie	Mean	1.0	0.89	0.85	1.0	0.99	0.95
	Median	1.0	0.96	0.86	1.0	0.99	0.95
Credit Card	Mean	0.84	0.64	0.79	0.94	0.02	0.81
	Median	0.89	0.71	0.81	0.98	0.21	0.83

- Agreement with SHAP values regarding the order of importance of features, (**high Kendall coefficient**).
- Very similar attribution scores (**R^2-score**).

Conclusion

- Eliminating the random step in Kernel SHAP leads to explanation stability by construction, and maintains high fidelity.
- Removing randomness still maintains high fidelity.
- Using only stratum 1 achieves complete stability and good fidelity.

Future Work

- Understand the properties of black-box models that guarantee good approximations of the SHAP values.
- Explore the relationship between the budget and the complexity of the target black-box we aim to explain.

References

[1] Lundberg, S. M.; and Lee, S.-I. 2017. **A unified approach to interpreting model predictions**. NIPS
[2] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. **" Why should i trust you?" Explaining the predictions of any classifier**. In Proceedings of the 22nd ACM SIGKDD
[3] Shrikumar, A. ; Greenside, P. ; and Kundaje, A. 2017. **Learning important features through propagating activation differences**. In International conference on machine learning, PMLR
[4] Shapley, L. S.; et al. 1953. **A Value for n-Person Games**. In Kuhn, H. W.; and Tucker, A. W., eds., Contributions to the Theory of Games (AM-28), Volume II, 307–318.
[5] https://github.com/gwladyskelodjou/st-shap. **Contact** : gwladys.kelodjou@irisa.fr
[6] Ruiz, L. M.; Valenciano, F.; and Zarzuelo, J. M. 1998. **The family of least square values for transferable utility games**. Games and Economic Behavior
[7] Condevaux, C.; Harispe, S.; and Mussard, S. 2022. **Fair and Efficient Alternatives to Shapley-based Attribution Methods**. ECML PKDD