Enhancing Model Interpretability through Stochastic Dimensionality Reduction and KernelSHAP

Conceptual Clarification and Practical Approaches*

Niyang Bai†

March 12, 2024

Abstract

This thesis proposes an innovative approach to enhance the robustness of KernelSHAP explanations by integrating stochastic dimensionality reduction techniques, specifically focusing on high-dimensional data like images. By preprocessing images into super-pixels and averaging KernelSHAP results over repeated stochastic feature groupings, we aim to establish a more reliable explanation method for supervised learning models. This research contributes to the field of Explainable Artificial Intelligence (XAI) by offering a scalable and interpretable method for model explanation in high-dimensional settings.

1 Introduction

Shapley Additive Explanation (SHAP) values are at the forefront of Explainable Artificial Intelligence (XAI), providing local explanations for predictions made by any supervised learning model. KernelSHAP, an efficient SHAP values approximation method, faces challenges in high-dimensional spaces due to computational infeasibility. This thesis investigates the improvement of KernelSHAP predictions through stochastic dimensionality reduction, aiming to enhance explanation robustness and interpretability in high-dimensional datasets, such as images.

2 Literature Review

SHAP values, introduced by Lundberg and Lee (2017), utilize cooperative game theory to interpret machine learning predictions, offering a model-agnostic framework that unifies feature importance measures. The work of Ribeiro, Singh, and Guestrin (2016) on LIME emphasized local interpretability and model-agnostic

[*] Please find the source code here: github.com/niyangbai/StochasticSHAP.git

[†] Please find the contact information here: niyang.bai@fau.de

explanations, inspiring enhancements in SHAP methodologies, including KernelSHAP for complex models. Wang et al. (2017) discussed superpixel segmentation for dimensionality reduction in image data, aiding feature analysis. This thesis proposes integrating these insights to improve KernelSHAP's efficiency and interpretability for high-dimensional data, incorporating superpixel segmentation for tractable feature spaces and building on the local explanation and theoretical foundation of SHAP values to develop a robust, computationally feasible explanation method for complex models.

3 Problem Statement

Despite the widespread utility and acceptance of KernelSHAP as a pivotal tool in the domain of Explainable Artificial Intelligence (XAI), its applicability and performance significantly degrade in high-dimensional settings, particularly in contexts involving complex data types such as images and large-scale datasets. This degradation primarily stems from computational infeasibility, as the exhaustive computation of SHAP values for each possible feature subset becomes exponentially difficult with increasing dimensionality. Consequently, this limitation not only hampers the scalability of KernelSHAP but also affects the accuracy and reliability of the explanations it generates, thus undermining the model's interpretability and the trust that users can place in its predictions.

In response to these challenges, this work seeks to innovatively address the computational bottleneck of KernelSHAP in high-dimensional spaces through the incorporation of stochastic dimensionality reduction techniques. By intelligently reducing the dimensionality of the input feature space in a stochastic manner, this approach aims to preserve the essential characteristics and variabilities of the data while significantly reducing the computational complexity involved in calculating SHAP values. This methodological enhancement is anticipated to not only improve the performance and scalability of KernelSHAP across various complex datasets but also enhance the robustness and reliability of the explanations generated, thereby bridging a critical gap in the literature on explainable AI and facilitating a deeper understanding of model predictions in high-dimensional applications.

4 Objectives

The primary objectives of this thesis are detailed below, focusing on the development and evaluation of an innovative approach to computational efficiency and interpretability in SHAP explanations:

1. To develop an algorithm that systematically applies stochastic dimensionality reduction techniques for computing KernelSHAP values. This algorithm aims to efficiently preprocess high-dimensional data, such as images or complex datasets, into a reduced feature space that retains the

essential information necessary for generating accurate and interpretable SHAP explanations.

2. To evaluate the effectiveness of this novel approach in enhancing the robustness and interpretability of SHAP explanations across a variety of model types and datasets. This includes a thorough assessment of the algorithm's performance in reducing computational complexity, its ability to maintain or improve the accuracy of SHAP values, and its impact on the overall interpretability of the explanations. The evaluation will be conducted through extensive experiments on benchmark datasets in different domains, including but not limited to image classification, text analysis, and genomic data interpretation, covering a broad spectrum of supervised learning models.

3. To compare the proposed method with existing dimensionality reduction techniques and SHAP computation strategies, thereby highlighting its advantages and potential limitations. This comparative analysis aims to provide a comprehensive understanding of how stochastic dimensionality reduction can be optimally utilized in the context of explainable AI, offering insights into best practices and future research directions in the field.

5 Methodology

This research employs advanced stochastic processes for dimensionality reduction, focusing on the transformation of high-dimensional image data into a manageable form through super-pixel segmentation. This method aims to preserve the structural integrity and essential characteristics of the images while reducing computational complexity. The methodology encompasses the development and implementation of a novel algorithm, Stochastic KernelSHAP with Super-pixels, which leverages stochastic dimensionality reduction to improve the robustness and interpretability of SHAP explanations.

5.1 Super-pixel Segmentation

Super-pixels are aggregated pixel clusters in an image that are more meaningful and computationally efficient to analyze than individual pixels. By grouping pixels based on color similarity and spatial proximity, super-pixels retain much of the original image's contextual information with significantly reduced complexity. This preprocessing step is crucial for efficiently applying KernelSHAP to high-dimensional data such as images.

5.2 Stochastic Dimensionality Reduction

The stochastic dimensionality reduction process involves randomly selecting subsets of super-pixels in each iteration to compute their contribution to the model's prediction. This randomness introduces variability, enabling a more robust estimation of feature importance by capturing a wide range of interactions among super-pixels.

5.3 KernelSHAP Computation

KernelSHAP, a model-agnostic method, will be applied to the reduced-dimensionality data to compute SHAP values. These values quantify the contribution of each super-pixel (or group of super-pixels in the reduced space) to the prediction, providing insights into the model's decision-making process.

Algorithm 1 Stochastic KernelSHAP with Super-pixels

```
Input: Image dataset $D$, Number of iterations $N$
Output: Robust SHAP values
for $i = 1$ to $N$ do
    Preprocess $D_i$ into super-pixels using a segmentation algorithm
    Randomly select a subset of super-pixels from $D_i$ to form a reduced dataset $D_{i,reduced}$
    Compute KernelSHAP on $D_{i,reduced}$ to obtain SHAP values for the selected super-pixels
    Store the computed SHAP values in an aggregation matrix
end for
Normalize and average the SHAP values stored in the aggregation matrix across all iterations
return Averaged SHAP values for super-pixels across $N$ iterations
```

The effectiveness of the proposed method will be assessed through rigorous comparisons with traditional KernelSHAP approaches, with a focus on improvements in accuracy, computational efficiency, and interpretability of explanations. By systematically analyzing the impact of stochastic dimensionality reduction on the robustness of SHAP explanations, this research aims to contribute significantly to the field of Explainable AI.

5.4 Evaluation Criteria

The evaluation of the proposed methodology will be conducted based on the following criteria:

- **Accuracy:** The fidelity of SHAP values in representing the true contribution of features (super-pixels) to the model's predictions.

- **Computational Efficiency:** The reduction in computational time and resources compared to conventional high-dimensional SHAP value computation methods.

- **Interpretability:** The clarity and usefulness of the generated explanations in providing insights into the model's decision-making process.

By addressing these criteria, the research will provide a comprehensive understanding of the proposed method's effectiveness in enhancing the robustness and utility of KernelSHAP explanations.

6 Expected Outcomes

The anticipated outcome is a more robust and computationally viable method for generating SHAP explanations in high-dimensional datasets, contributing significantly to the field of XAI.

7 Novelty of the Thesis

This thesis introduces a novel approach to explainability in AI by incorporating stochastic dimensionality reduction into the computation of KernelSHAP values. Unlike previous methods, this approach aims to provide a scalable and robust mechanism for generating explanations in high-dimensional feature spaces, such as those encountered in image classification tasks. The use of repeated stochastic preprocessing and aggregation of SHAP values across multiple iterations represents an innovative step towards more accurate and reliable model explanations.

8 Conclusion

In conclusion, this thesis proposes a groundbreaking methodology for enhancing the robustness, computational efficiency, and interpretability of SHAP explanations in high-dimensional datasets. By innovatively applying stochastic dimensionality reduction and super-pixel segmentation, this research not only tackles existing limitations in the field of XAI but also opens new avenues for further exploration and development. The expected outcomes underscore the potential of this approach to significantly advance the interpretability of complex models, thus contributing to the creation of more transparent, understandable, and trustworthy AI systems.

References

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Wang, Murong et al. (2017). “Superpixel segmentation: A benchmark”. In: *Signal Processing: Image Communication* 56, pp. 28–39.