Enhancing Model Interpretability through Sparse Data and KernelSHAP

Niyang Bai

July 28, 2024

Abstract

This thesis proposes an innovative approach to enhance the robustness of KernelSHAP explanations by integrating techniques specifically designed for sparse data. By preprocessing data through one-hot encoding and applying various stochastic methods, we aim to establish a more reliable explanation method for supervised learning models dealing with high-dimensional sparse data. This research contributes to the field of Explainable Artificial Intelligence (XAI) by offering scalable and interpretable methods for model explanation in sparse data 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 and sparse data spaces due to computational infeasibility. This thesis investigates the improvement of KernelSHAP predictions through stochastic methods tailored for sparse data, aiming to enhance explanation robustness and interpretability.

2 Literature Review

SHAP values, introduced by Lundberg and Lee (Lundberg & 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 (Ribeiro, Singh, & Guestrin, 2016) on LIME emphasized local interpretability and model-agnostic explanations, inspiring enhancements in SHAP methodologies. This thesis proposes integrating these insights to improve KernelSHAP's efficiency and interpretability for high-dimensional sparse data.

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 and sparse settings. 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 and sparsity. Consequently, this limitation hampers the scalability of KernelSHAP, affecting the accuracy and reliability of the explanations it generates.

4 Objectives

The primary objectives of this thesis are:

1. **Algorithm Development:** To develop algorithms that systematically apply stochastic methods tailored for sparse data when computing KernelSHAP values. This includes handling one-hot encoded attributes efficiently.
2. **Effectiveness Evaluation:** To evaluate the effectiveness of these methods in enhancing the robustness and interpretability of SHAP explanations across various model types and sparse datasets.
3. **Comparative Analysis:** To compare the proposed methods with existing SHAP computation strategies, highlighting their advantages and potential limitations.

5 Methodology

5.1 Naive SHAP
Benchmark

As a benchmark, we utilize the Naive SHAP computation formula:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

This formula considers all possible subsets S of features excluding i, and calculates the average marginal contribution of feature i across all subsets.

5.2 Stochastic Group and KernelSHAP (SG-KernelSHAP)
Selecting Only Non-Zeros

We focus on selecting only non-zero elements in the sparse data:

Mathematical Proof of Equivalence:

Algorithm 1 SG-KernelSHAP with Non-Zero Selection

1: Identify non-zero features in the one-hot encoded data.
2: Add one-hot encoded attributes to the original attributes.
3: Prove that zeros are not relevant by demonstrating that including them does not change the SHAP value contribution.
4: Apply KernelSHAP to this reduced feature set.

For a feature i with value zero, including it in any subset S does not change the function output $f(S)$. Thus, the marginal contribution $f(S \cup \{i\}) - f(S) = 0$. Consequently, zero-valued features do not affect the sum of contributions, proving the reduced set with only non-zero features maintains equivalence with the Naive SHAP.

Let's formalize this:

Identify non-zero features: Let $Z = \{j \mid x_j = 0\}$ be the set of zero features.

For any subset $S \subseteq N \setminus \{i\}$:

$$f(S \cup \{i\}) - f(S) = \begin{cases} f(S) - f(S) = 0 & \text{if } i \in Z \\ f(S \cup \{i\}) - f(S) & \text{if } i \notin Z \end{cases}$$

Therefore, the SHAP value for zero features:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[0] = 0$$

Hence, considering only non-zero features maintains the Naive SHAP formula for relevant features.

Computational Efficiency: This method reduces the number of features considered in each SHAP computation. Computational complexity is reduced from $O(2^n)$ to $O(2^k)$, where k is the number of non-zero features.

5.3 PCA and SG-KernelSHAP
Dimensionality Reduction through PCA

Algorithm 2 PCA and SG-KernelSHAP

1: Apply PCA to select top n principal components from the data.
2: Use these components for stochastic grouping and KernelSHAP computation.
3: Map the SHAP values back to the original feature space.
4: Add one-hot encoded attributes to the original attributes.

Mathematical Proof of Equivalence:

PCA Transformation: Let X be the original data matrix and P be the matrix of principal components. The transformed data $X' = XP$.

KernelSHAP on Transformed Data: Compute KernelSHAP values on X'. Let ϕ'_i be the SHAP value in the transformed space.

Inverse Mapping: Map ϕ'_i back to the original space using $\phi_i = P\phi'_i$.

The contributions in the principal component space reflect an approximation of those in the original space. PCA reduces dimensions while retaining the variance, which is critical for maintaining the integrity of SHAP values.

Since PCA only approximates the original features, the SHAP values in the principal component space are an approximation:

$$\phi_i \approx \sum_{S' \subseteq P} \frac{|S'|!(|P| - |S'| - 1)!}{|P|!} [f(S' \cup \{i'\}) - f(S')] \cdot P^{-1}$$

While not exact, if the principal components capture most of the variance, this approximation can be highly accurate:

$$\phi_i \approx \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

Computational Efficiency: This method reduces dimensionality significantly before SHAP computation. Complexity is reduced due to fewer principal components ($O(2^m)$, where m is the number of principal components).

5.4 SG-KernelSHAP with Grouped Features *Grouping Features*

Algorithm 3 SG-KernelSHAP with Grouped Features

1: Group features into meaningful sets based on domain knowledge or feature similarity.
2: Apply stochastic grouping to these feature sets and compute KernelSHAP values.
3: Combine the SHAP values for grouped features and map them back to individual features if needed.

Mathematical Proof of Equivalence:

Feature Grouping: Let G be a set of grouped features. The SHAP value for a group G_i is computed as:

$$\phi_{G_i} = \sum_{S \subseteq N \setminus \{G_i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{G_i\}) - f(S)]$$

Distribute Contribution: Distribute the group's SHAP value to individual features within the group:

$$\phi_{i \in G} = \frac{\phi_G}{|G|}$$

The distribution maintains the total contribution equivalence by ensuring the sum of SHAP values of individual features in the group equals the group's SHAP value:

$$\sum_{i \in G} \phi_i = \phi_G = \sum_{S \subseteq N \setminus \{G\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{G\}) - f(S)]$$

Computational Efficiency: This method groups features to reduce the effective dimensionality. Computational complexity is reduced based on the number of groups rather than individual features.

6 Expected Outcomes

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

7 Novelty of the Thesis

This thesis introduces novel approaches to explainability in AI by incorporating stochastic methods tailored for sparse data into the computation of KernelSHAP values. These methods provide scalable and robust mechanisms for generating explanations, representing an innovative step towards more accurate and reliable model explanations.

8 Conclusion

In conclusion, this thesis proposes groundbreaking methodologies for enhancing the robustness, computational efficiency, and interpretability of SHAP explanations in sparse data settings. By innovatively applying stochastic methods and dimensionality reduction techniques, this research tackles existing limitations in XAI, opening new avenues for further exploration and development. The expected outcomes underscore the potential of these approaches to significantly advance the interpretability of complex models, contributing to the creation of more transparent, understandable, and trustworthy AI systems.

References

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (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).