Enhancing Model Interpretability with Sparse Data

Niyang Bai
niyang.bai@fau.de

July 28, 2024

Abstract

This thesis proposes an innovative approach to enhance the robustness of KernelSHAP explanations by integrating techniques tailored for sparse data, specifically focusing on one-hot encoded features. By preprocessing sparse data and leveraging stochastic groupings, PCA, and feature group setups, 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 and sparse settings.

1 Introduction

Shapley Additive Explanation (SHAP) values are a pivotal tool in Explainable Artificial Intelligence (XAI), providing local explanations for predictions made by any supervised learning model. KernelSHAP, an efficient method to approximate SHAP values, faces significant challenges in high-dimensional and sparse spaces due to computational infeasibility. The essence of this thesis lies in investigating the improvement of KernelSHAP predictions by leveraging stochastic feature grouping, Principal Component Analysis (PCA), and strategic feature setups. These methodologies aim to enhance the robustness and interpretability of SHAP explanations when dealing with sparse datasets, such as those with predominantly one-hot encoded features.

2 Literature Review

The concept of SHAP values, introduced by Lundberg and Lee (2017), utilizes cooperative game theory to provide a unified framework for interpreting machine learning predictions, ensuring consistency and local accuracy. KernelSHAP, designed for complex models, faces computational challenges in high-dimensional and sparse data settings.

Ribeiro et al. (2016) emphasized local interpretability in their work on LIME, which inspired enhancements in SHAP methodologies. Lundberg et al. (2020) further advanced SHAP with TreeSHAP for tree-based models, highlighting scalability improvements.

Dimensionality reduction techniques, as discussed by Jolliffe (2002) through Principal Component Analysis (PCA), reduce computational complexity while retaining essential data characteristics. Hastie et al. (2009) explored stochastic processes, and Goodfellow et al. (2016) discussed deep learning optimization, providing foundational methods to enhance SHAP computations.

Murphy (2012) and Ng (2004) address challenges in handling sparse data, emphasizing the importance of feature selection for computational efficiency. These works collectively underscore the necessity of efficient SHAP computation methods, guiding this thesis to develop innovative approaches for robust and interpretable SHAP explanations in sparse datasets.

3 Problem Statement

KernelSHAP, while widely accepted and utilized in XAI, suffers from degraded performance in sparse settings due to the heavy computational burden of evaluating numerous feature subsets. This limitation hampers the scalability of KernelSHAP and affects the accuracy and reliability of the explanations it generates, ultimately undermining model interpretability and user trust. The core challenge addressed in this thesis is to develop methods that can efficiently handle the computational complexity of SHAP value computation in sparse data, preserving essential characteristics and improving the interpretability of the explanations provided.

4 Objectives

The primary objectives of this thesis are to:

- **Develop Algorithms**: Create innovative algorithms that apply stochastic feature groupings, PCA, and grouped feature setups for sparse data to compute KernelSHAP values. These algorithms aim to preprocess sparse datasets to retain essential information necessary for accurate and interpretable SHAP explanations while significantly reducing computational complexity.
- **Evaluate Effectiveness**: Assess the effectiveness of these approaches in enhancing the robustness and interpretability of SHAP explanations across various model types and datasets. This includes focusing on:
  - Maintaining or improving the accuracy of SHAP values.
  - Reducing computational time and resources.
  - Enhancing the overall interpretability of the explanations.

- **Comparative Analysis**: Conduct a comparative analysis with traditional KernelSHAP strategies to highlight the advantages and potential limitations of the proposed methods. This will provide a comprehensive understanding of how stochastic dimensionality reduction and feature grouping can be optimally utilized in XAI.

5 Methodology and Comparison

5.1 Naive SHAP (Benchmark)

The naive SHAP method provides a comprehensive way to explain a machine learning model's predictions by distributing the prediction among the features based on their contribution. This is done using Shapley values from cooperative game theory, which considers all possible combinations of feature subsets to ensure that each feature's contribution is fairly evaluated. This method is computationally intensive because it requires evaluating the model on every possible subset of features.

Algorithm 1 Naive SHAP

Input: Set of all features $X = \{x_1, x_2, \ldots, x_n\}$
for each feature x_i **do**
    for each subset $S \subseteq N \setminus \{i\}$ **do**
        Compute the marginal contribution $v(S \cup \{i\}) - v(S)$
        Weight the marginal contribution by $\frac{|S|!(|N|-|S|-1)!}{|N|!}$
    end for
    Aggregate the weighted marginal contributions to obtain ϕ_i
end for
Output: SHAP values ϕ_i for each feature x_i

Complexity: The computational complexity of naive SHAP is $O(2^n)$ since it requires evaluating the model on every possible subset of n features.

5.2 Stochastic Group and KernelSHAP - Selecting Only Non-Zeros

This algorithm aims to simplify the SHAP value computation for sparse data by focusing only on the non-zero features. Since zero entries in one-hot encoded data do not contribute to feature interactions, excluding them from the SHAP computation reduces computational complexity without affecting the results.

Algorithm 2 Stochastic Group and KernelSHAP (Selecting Non-Zeros)

Input: Set of features $X = \{x_1, x_2, \ldots, x_n\}$
Identify non-zero one-hot encoded features $Z \subseteq X$
for each feature $x_i \in Z$ **do**
    for each subset $S \subseteq Z \setminus \{i\}$ **do**
        Compute the marginal contribution $v(S \cup \{i\}) - v(S)$
        Weight the marginal contribution by $\frac{|S|!(|Z|-|S|-1)!}{|Z|!}$
    end for
    Aggregate the weighted marginal contributions to obtain ϕ_i
end for
Aggregate the SHAP values over multiple stochastic selections
Output: SHAP values ϕ_i for each feature x_i

Mathematical Proof and Justification:

Consider the set of features $X = \{x_1, x_2, \ldots, x_n\}$ where some features are one-hot encoded and hence sparse. Let $Z \subseteq X$ be the set of non-zero features, and let Z_0 be the set of features that are zero in the instance being explained.

For KernelSHAP, we select a subset S of Z and compute the SHAP value:

$$\phi_i = \sum_{S \subseteq Z \setminus \{i\}} \frac{|S|!(|Z| - |S| - 1)!}{|Z|!} (v(S \cup \{i\}) - v(S))$$

Since zeros in one-hot encoded features indicate the absence of categorical attributes, their contribution to the value function $v(S)$ is effectively neutral. Specifically, for any $S \subseteq Z$:

$$v(S \cup Z_0) = v(S)$$

This implies that the value of the function does not change with the inclusion of zero-valued features, meaning their marginal contribution is zero:

$$v(S \cup \{i\}) - v(S) = 0 \quad \text{if} \quad i \in Z_0$$

Therefore, the SHAP value for a zero feature $i \in Z_0$ is:

$$\begin{aligned}
\phi_i &= \sum_{S \subseteq Z \setminus \{i\}} \frac{|S|!(|Z| - |S| - 1)!}{|Z|!} \cdot (v(S \cup \{i\}) - v(S)) \\
&= \sum_{S \subseteq Z \setminus \{i\}} \frac{|S|!(|Z| - |S| - 1)!}{|Z|!} \cdot 0 \\
&= 0
\end{aligned}$$

Thus, zero-valued features do not contribute to the SHAP value, confirming that they are not relevant in the computation. The expected value of the SHAP

value remains the same whether zero features are included or excluded:

$$\mathbb{E}[\phi_i^{\text{Stochastic}}] = \mathbb{E}\left[\sum_{S \subseteq Z \setminus \{i\}} \frac{|S|!(|Z| - |S| - 1)!}{|Z|!}(v(S \cup \{i\}) - v(S))\right]$$
$$= \phi_i$$

Hence, the expected value of the SHAP value using stochastic selection of non-zero features is equivalent to the naive SHAP value.

Complexity: Let m be the number of non-zero features in a particular instance. The computational complexity of this method is $O(2^m)$, which is significantly lower than $O(2^n)$ if $m \ll n$.

5.3 PCA, Stochastic Grouping, and KernelSHAP

This algorithm integrates Principal Component Analysis (PCA) to reduce the dimensionality of the original features while retaining the most significant components. The reduced feature set, combined with the non-zero one-hot encoded features, allows for efficient SHAP value computation. The results are then mapped back to the original feature space.

Algorithm 3 PCA, Stochastic Grouping, and KernelSHAP

Input: Set of original features X
Apply PCA to the original features and select the top k principal components
Combine the selected principal components with non-zero one-hot encoded features Z
for each feature $x_i \in X' = \{\text{top } k \text{ principal components}\} \cup Z$ **do**
    for each subset $S' \subseteq X' \setminus \{i\}$ **do**
        Compute the marginal contribution $v(S' \cup \{i\}) - v(S')$
        Weight the marginal contribution by $\frac{|S'|!(|X'|-|S'|-1)!}{|X'|!}$
    end for
    Aggregate the weighted marginal contributions to obtain ϕ_i'
end for
Map the SHAP values back to the original feature space
Output: SHAP values ϕ_i for each feature x_i

Mathematical Proof and Justification:

Let X be the original feature set and X' be the transformed feature set after applying PCA, where X' consists of the top k principal components.

The SHAP value is computed on X':

$$\phi_i' = \sum_{S' \subseteq X' \setminus \{i'\}} \frac{|S'|!(|X'| - |S'| - 1)!}{|X'|!}(v(S' \cup \{i'\}) - v(S'))$$

Mapping back to the original space, the SHAP values ϕ_i can be approximated by:

$$\phi_i \approx \sum_{j=1}^{k} \alpha_{ij} \phi'_j$$

where α_{ij} are the coefficients that map the principal components back to the original features.

The expected value of the SHAP value using PCA and stochastic grouping is:

$$\mathbb{E}[\phi_i^{\text{PCA}}] = \mathbb{E}\left[\sum_{j=1}^{k} \alpha_{ij} \sum_{S' \subseteq X' \setminus \{j\}} \frac{|S'|!(|X'| - |S'| - 1)!}{|X'|!} (v(S' \cup \{j\}) - v(S'))\right]$$

Given the linearity of expectation and the properties of PCA preserving variance and interactions:

$$\mathbb{E}[\phi_i^{\text{PCA}}] = \phi_i$$

Thus, the expected value of the SHAP value using PCA and stochastic grouping is equivalent to the naive SHAP value.

Complexity:

- PCA transformation has a complexity of $O(n^3)$ for computing the principal components.
- After PCA, the complexity of KernelSHAP on k components and m non-zero features is $O(2^k \cdot 2^m)$. If k and m are significantly smaller than n, this complexity is much lower than $O(2^n)$.

5.4 Stochastic Grouping and KernelSHAP with Grouped Feature Setup

This algorithm groups one-hot encoded features based on their original categorical attributes, allowing for a more structured approach to feature selection. By applying stochastic grouping within these groups, the algorithm maintains the integrity of feature interactions and reduces dimensionality.

Algorithm 4 Stochastic Grouping and KernelSHAP with Grouped Feature Setup

Input: Set of one-hot encoded features grouped by original categorical attributes $G = \{G_1, G_2, \ldots, G_m\}$
for each group $G_j \in G$ **do**
  for each feature $x_i \in G_j$ **do**
    for each subset $S_j \subseteq G_j \setminus \{i\}$ **do**
      Compute the marginal contribution $v(S_j \cup \{i\}) - v(S_j)$
      Weight the marginal contribution by $\frac{|S_j|!(|G_j|-|S_j|-1)!}{|G_j|!}$
    end for
    Aggregate the weighted marginal contributions to obtain ϕ_{ij}
  end for
end for
Aggregate the SHAP values across all groups to obtain ϕ_i
Output: SHAP values ϕ_i for each feature x_i

Mathematical Proof and Justification:

Let $G = \{G_1, G_2, \ldots, G_m\}$ be the groups of features, where each group G_j contains related one-hot encoded features.

For KernelSHAP with grouped features, we compute the SHAP value within each group G_j:

$$\phi_{ij} = \sum_{S_j \subseteq G_j \setminus \{i\}} \frac{|S_j|!(|G_j| - |S_j| - 1)!}{|G_j|!} (v(S_j \cup \{i\}) - v(S_j))$$

The overall SHAP value for feature x_i considering the grouped setup is:

$$\phi_i = \sum_{j=1}^{m} \phi_{ij}$$

Since the grouping preserves the interaction patterns and contributions of features, the expected value is:

$$\mathbb{E}[\phi_i^{\text{Grouped}}] = \mathbb{E}\left[\sum_{j=1}^{m} \sum_{S_j \subseteq G_j \setminus \{i\}} \frac{|S_j|!(|G_j| - |S_j| - 1)!}{|G_j|!} (v(S_j \cup \{i\}) - v(S_j))\right]$$

Again, by the linearity of expectation and the grouping preserving feature interactions:

$$\mathbb{E}[\phi_i^{\text{Grouped}}] = \phi_i$$

Thus, the expected value of the SHAP value using the grouped feature setup is equivalent to the naive SHAP value.

Complexity:

- Let g be the number of groups and h be the average number of features per group.

- The complexity of KernelSHAP for each group is $O(2^h)$.
- Since there are g groups, the total complexity is $O(g \cdot 2^h)$. If h is significantly smaller than n, this complexity is much lower than $O(2^n)$.

5.5 Comparison of Complexities

- **Naive SHAP:** $O(2^n)$
- **Stochastic Group and KernelSHAP:** $O(2^m)$ (where m is the number of non-zero features)
- **PCA, Stochastic Grouping, and KernelSHAP:** $O(n^3 + 2^k \cdot 2^m)$ (where k is the number of principal components and m is the number of non-zero features)
- **Stochastic Grouping and KernelSHAP with Grouped Feature Setup:** $O(g \cdot 2^h)$ (where g is the number of groups and h is the average number of features per group)

Each method significantly reduces the complexity compared to the naive SHAP approach, making SHAP computations feasible for high-dimensional and sparse data.

6 Evaluation Criteria

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

- **Accuracy**: Measure the fidelity of the SHAP values in representing the true contribution of features to the model's predictions.
- **Computational Efficiency**: Assess the reduction in computational time and resources compared to conventional high-dimensional SHAP value computation methods.
- **Interpretability**: Evaluate 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.

7 Expected Outcomes

The anticipated outcome of this research is to develop more robust and computationally viable methods for generating SHAP explanations in sparse data settings. These methods are expected to significantly enhance the interpretability of complex models by providing scalable and accurate explanations that can be understood and trusted by users. The innovative approaches proposed in this thesis aim to contribute substantially to the field of XAI by addressing current limitations in KernelSHAP and opening new avenues for further exploration and development.

8 Novelty of the Thesis

This thesis introduces a novel approach to explainability in AI by incorporating stochastic dimensionality reduction and strategic feature setups into the computation of KernelSHAP values for sparse data. Unlike previous methods, this approach focuses on providing scalable and robust mechanisms for generating explanations in high-dimensional and sparse feature spaces. The use of repeated stochastic preprocessing, PCA, and feature grouping represents an innovative step towards more accurate and reliable model explanations, offering significant advancements in the field of XAI.

9 Conclusion

In conclusion, this thesis proposes groundbreaking methodologies to enhance the robustness, computational efficiency, and interpretability of SHAP explanations in sparse datasets. By innovatively applying stochastic dimensionality reduction, PCA, and feature grouping, this research addresses existing limitations in XAI and provides scalable solutions for interpreting complex models. The expected outcomes underscore the potential of these approaches to significantly advance the interpretability of AI systems, contributing to the creation of more transparent, understandable, and trustworthy models.

References

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Jolliffe, I. T. (2002). Principal component analysis and factor analysis. *Principal component analysis*, pages 150–166.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

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

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Ng, A. Y. (2004). Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78.

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 international conference on knowledge discovery and data mining*, pages 1135–1144.