Trustworthy Clustering: A Interpretable Clustering Framework with Hierarchical Oblique Decision Boundaries

Author information scrubbed for double-blind reviewing

institute {email}

Abstract. The field of Explainable AI (XAI) has largely focused on interpretability in classification and regression, while cluster analysis, a vital unsupervised machine learning technique, has been relatively neglected. As an exploratory technique, the interpretability and explainability of modern clustering models present significant challenges. This lack of transparency impedes trust and comprehension of the models' decision-making processes. In this paper, we address this gap by enhancing the interpretability of clustering results using a hierarchical structure representation. Our approach employs oblique decision trees, supported by SHAP (SHapley Additive exPlanations) values to analyze influential features and identify separation hyperplanes for constructing the trees. Unlike traditional axis-aligned trees, oblique decision trees provide a more accurate interpretation of high-dimensional data, maintaining a clear and interpretable structure. This method not only improves transparency but also fosters trust in hybrid decision-making systems, offering reliable and comprehensible explanations for clustering outcomes.

Keywords: Explainable AI · Trustworthy machine learning · Clustering · Boundary detection · Oblique decision trees.

1 Introduction

Clustering, a fundamental unsupervised machine learning technique, plays a crucial role in various domains such as marketing, healthcare, and bioinformatics. Despite its widespread applications, the interpretability and explainability of modern clustering models remain a significant challenge, particularly in the context of trustworthy decision-making systems. Traditional clustering algorithms, while effective in grouping similar data points, often operate as "black boxes," lacking transparency in their decision-making processes. This opacity hinders trust in clustering outcomes, especially in high-stakes domains where explainability is paramount. While Explainable Artificial Intelligence (XAI) has made significant progress in developing interpretable models for classification and regression tasks, cluster analysis has received comparatively less attention.

We believe that addressing this interpretability gap in clustering is crucial for advancing trustworthy hybrid decision-making systems. Our research proposes a novel approach to elucidate cluster boundaries using a hierarchical structure

representation and oblique decision trees. By leveraging SHAP (SHapley Additive exPlanations) values, we analyze influential features and incorporate them into the construction of oblique trees, offering a more accurate interpretation of high-dimensional data compared to traditional methods.

This paper aims to demonstrate how our proposed method enhances the interpretability of clustering models, contributing to the development of more transparent and reliable decision-making systems. Our work has significant implications for improving the adoption of clustering techniques in domains where explainable AI is essential.

2 Background

2.1 Cluster analysis and boundary detection

Given a dataset $X = \{x_1, x_2, ..., x_n\}$, where each data point x_i is a d-dimensional vector representing the features of an observation. The goal is to partition this dataset into K clusters, $C = \{C_1, C_2, ..., C_K\}$, such that the clustering cost J such as the within-cluster sum of squares (WCSS) is minimized (shown as equation 1.).

$$J(C, \mu) = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \tag{1}$$

where K is the number of clusters, C_k is the set of data points in the k-th cluster. μ_k denotes the centroid of the k-th cluster, calculated as $(1/|C_k|) \sum_{x \in C_k} x$ for K-Means algorithm.

To explicitly identify a cluster, one of the most important aspects of the analysis is to find its boundary, particularly when dealing with complex or irregularly shaped clusters. After obtaining the cluster assignments and centroids from the K-Means algorithm, the boundaries between clusters can be estimated based on certain similarity measures such as pairwise distances or underlying density functions. For example, distance-based boundary detection methods rely on the idea that points located near the boundaries between clusters should have similar distances to two or more cluster centroids [5]. These methods aim to identify the boundary regions by finding the set of points that are approximately equidistant from multiple centroids. Specifically, for each data point x, find the closest and second-closest centroids, denoted as μ_{c1}^x and μ_{c2}^x, respectively. If the difference between the distances to these two centroids is less than a predefined threshold δ, then x is considered a potential boundary point: $|dist(x, \mu_{c1}^x) - dist(x, \mu_{c1}^x)| < \delta$. The threshold δ determines the width of the boundary region. A smaller value of δ will result in a narrower boundary region, while a larger value will produce a broader boundary region.

2.2 Oblique decision trees

Traditional decision trees, such as CART (Classification and Regression Trees) and C4.5, use axis-parallel splits, which can be limiting for data with complex re-

lationships between features, potentially resulting in suboptimal splits. Oblique decision trees, on the other hand, are a variant of traditional axis-parallel decision trees. Their splits consist of hyperplanes that are not constrained to being parallel to the feature axes. This allows for more flexible and accurate decision boundaries, especially in scenarios where class determination relies on a combination of features simultaneously.

To build such decision trees that utilize oblique splits, Sreerama *et al.* introduced the OC1 algorithm [3][4]. Initially, at each internal node of the decision tree, the OC1 algorithm randomly generates a set of candidate oblique splitting hyperplanes. For each candidate oblique hyperplane, the system evaluates a splitting criterion (e.g., Gini impurity) to measure the quality of the split. The hyperplane that maximizes the splitting criterion is chosen as the best oblique split for that node. Particularly, OC1 primarily relies on deterministic hill climbing to search for appropriate oblique splits. It iteratively adjusts the coefficients of the linear function, essentially tilting and shifting the hyperplane in the data space so that the potential plane meets the chosen splitting criterion during evaluation. Once the best oblique split is determined, the instances are partitioned according to that split, and the process is recursively applied to the child nodes until a stopping criterion is met. Additionally, OC1 also applies a pruning technique to remove subtrees that do not significantly improve accuracy for preventing overfitting.

3 Arguments

The field of Explainable AI (XAI) has made significant strides in developing interpretable models for supervised learning tasks, but the domain of clustering, a crucial unsupervised learning technique, remains underexplored in terms of interpretability. This gap is particularly problematic given the widespread application of clustering across various domains and its importance in decision-making processes. Our research addresses three primary challenges in explainable clustering:

1. The difficulty in adapting supervised interpretability methods to unsupervised clustering tasks. While many existing explainability methods provide reliable and theoretically robust explanations, they are primarily designed for supervised learning tasks. Adapting these techniques or developing new ones specifically for unsupervised clustering tasks presents additional challenges.

2. The complexity of interpreting high-dimensional feature spaces commonly used in clustering algorithms. Many machine learning approaches advocate mapping original features to higher-dimensional spaces to uncover latent information, resulting in input instances being represented as high-dimensional vectors. Aligning these hidden characteristics with real-world features remains challenging. Additionally, the complexity of feature interactions increases, making it more difficult to provide human-understandable explanations.

3. The sensitivity of cluster boundaries to hyperparameters and initialization conditions, which complicates consistent interpretation. As an exploratory tech-

nique, the explanations are inherently influenced by the clustering process. Small changes in hyperparameter settings or randomized initialization can significantly alter the resulting cluster boundaries, leading to notably different cluster assignments. Consequently, this sensitivity can undermine trust in the explanations provided, making it difficult for users to rely on them for informed decision-making.

Addressing these challenges is crucial for advancing trustworthy hybrid decision-making systems. To this end, we propose a novel approach that combines the strengths of oblique decision trees with detected cluster boundary integration. This method offers several key advantages: First, it provides a more accurate interpretation of cluster boundaries compared to traditional axis-aligned decision trees. Second, it maintains an interpretable structure in the form of a top-down binary tree, making it accessible for human understanding. Finally, it effectively handles high-dimensional data by identifying critical features that significantly impact the clustering process. By developing this method, we contribute to the broader field of XAI and unsupervised learning, offering a framework for understanding complex, high-dimensional clustering outcomes in a manner that is both accurate and accessible to human interpretation.

4 Proposed Method

Our proposed method introduces an integrated framework that leverages the strengths of multiple techniques to enhance the interpretability of clustering models while maintaining their performance. The framework consists of three main stages:

Feature Filtering with Contribution Analysis: Gan et al. proposed using an adjusted SHAP approach to capture feature contributions in clustering [13]. We will apply this method to identify and select the most contributive features for each cluster. This step helps reduce dimensionality and focuses the interpretation on the most relevant aspects of the data. By utilizing the adjusted SHAP approach, we benefit from its game-theoretic framework for feature importance, which accounts for feature interactions and provides a robust measure of feature impact on cluster assignments.

Cluster Boundary Collection: Building upon the feature selection performed through SHAP analysis, our framework employs advanced boundary detection techniques ([6], [10], [11]) to identify multiple linear decision boundaries for each cluster. This crucial step aims to capture the complex shapes of cluster boundaries in high-dimensional spaces, providing a more accurate representation of the clustering structure while maintaining interpretability. By focusing on the most contributive features, we address the challenges of high-dimensional complexity and the interpretability-accuracy trade-off inherent in clustering tasks.

Our approach utilizes a combination of methods, including Support Vector Machines (SVM) with linear kernels, adaptive gradient-based approaches, and density-based boundary detection. We implement a one-vs-rest strategy with SVMs, leveraging their margins to assess boundary confidence. Additionally, we

employ iterative gradient-based methods to fine-tune boundary positions, ensuring they accurately reflect local data distributions around cluster interfaces. For clusters with non-uniform density, we incorporate density estimation techniques to identify potential boundaries in regions of rapid density change.

Oblique Decision Tree Construction: The construction of an oblique decision tree serves as the final and crucial stage in our framework for interpretable clustering. This stage translates the collected linear cluster boundaries into a hierarchical, human-readable structure that effectively captures the essence of the clustering model's decision-making process. Unlike traditional axis-aligned decision trees, our oblique trees can create splits using linear combinations of multiple features, allowing for a more accurate representation of non-orthogonal decision boundaries identified in the previous stages.

The construction process integrates the linear decision boundaries collected earlier as potential split candidates for tree nodes. We employ a top-down approach, where each level of the tree represents a more fine-grained partitioning of the data space. At each node, we select the most discriminative boundary or combination of boundaries to split the data, maximizing cluster separation. This process is optimized through techniques such as linear discriminant analysis or logistic regression to determine the best oblique splits.

To maintain interpretability, we implement pruning techniques to prevent overfitting and reduce tree complexity, balancing the trade-off between tree depth and accuracy. We also incorporate the feature importance information derived from the SHAP analysis, ensuring that splits near the root of the tree prioritize the most influential features. This integration of SHAP values enhances the tree's ability to provide meaningful explanations of cluster assignments.

The resulting oblique decision tree offers several advantages for cluster interpretation. It provides a hierarchical view of the cluster structure and allows for accurate capture of complex decision boundaries. Moreover, the trees naturally lend themselves to generating rule-based explanations for cluster assignments.

5 Conclusion

This paper presents a novel framework for enhancing the interpretability of clustering models, addressing a critical gap in the field of Explainable AI (XAI) for unsupervised learning tasks. Our framework addresses several key challenges in the field of explainable clustering. It mitigates the difficulty of adapting supervised interpretability methods to unsupervised tasks, handles the complexity of high-dimensional feature spaces, and provides consistent interpretations despite the sensitivity of clustering algorithms to initialization and hyperparameters. By doing so, it contributes significantly to the development of trustworthy hybrid decision-making systems, offering transparency, consistency, and adaptability across various clustering scenarios.

The primary contributions of our work are threefold. First, we involve the adapting SHAP approach to identify and prioritize the most influential features in clustering outcomes. This adaptation provides a robust foundation for feature

selection in an unsupervised context. Second, our boundary detection approach, which combines multiple techniques including SVMs and density-based methods, offers a sufficient representation of cluster boundaries that captures the complexity of real-world data distributions. Finally, the use of oblique decision trees as an interpretable model provides a flexible and accurate representation of cluster structures, bridging the gap between sophisticated clustering algorithms and human-understandable explanations.

Looking forward, this work opens up several avenues for future research. These include the exploration of more advanced oblique decision tree algorithms, to further enhance the accuracy and interpretability of our model. Then we may explore the integration of interactive visualization techniques to enhance user understanding instead of rule-based explanations. Next, we plan to conduct comprehensive comparative studies, evaluating our framework against a diverse array of interpretable clustering approaches across various dimensions, including robustness, and domain-specific applicability. Furthermore, a critical aspect of our future work involves the formulation of novel metrics for quantifying the quality of interpretable methods, addressing the current gap in standardized evaluation criteria for explainable clustering. Our consistent research aim is not only to refine our current framework but also to contribute to the broader advancement of interpretable and trustworthy machine learning in real-world contexts.

References

1. Montavon, Grégoire, et al. "Explaining the predictions of unsupervised learning models." International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. Cham: Springer International Publishing, 2020.
2. Brodley, Carla E., and Paul E. Utgoff. "Multivariate decision trees." Machine learning 19 (1995): 45-77.
3. Murthy, Sreerama K., et al. "OC1: A randomized algorithm for building oblique decision trees." Proceedings of AAAI. Vol. 93. Citeseer, 1993.
4. Murthy, Sreerama K., Simon Kasif, and Steven Salzberg. "A system for induction of oblique decision trees." Journal of artificial intelligence research 2 (1994): 1-32.
5. Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." IEEE Transactions on neural networks 16.3 (2005): 645-678.
6. Fukunaga, Keinosuke, and Larry Hostetler. "The estimation of the gradient of a density function, with applications in pattern recognition." IEEE Transactions on information theory 21.1 (1975): 32-40.
7. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
8. Thorndike, Robert. "Who belongs in the family?." Psychometrika 18.4 (1953): 267-276.
9. Khalique, Vijdan, and Hiroyuki Kitagawa. "BPF: an effective cluster boundary points detection technique." International Conference on Database and Expert Systems Applications. Cham: Springer International Publishing, 2022.
10. Cao, Xiaofeng, Baozhi Qiu, and Guandong Xu. "BorderShift: toward optimal MeanShift vector for cluster boundary detection in high-dimensional data." Pattern Analysis and Applications 22 (2019): 1015-1027.

11. Cao, Xiaofeng, et al. "Multidimensional balance-based cluster boundary detection for high-dimensional data." IEEE transactions on neural networks and learning systems 30.6 (2018): 1867-1880.
12. Peng, Xi, et al. "XAI beyond classification: Interpretable neural clustering." Journal of Machine Learning Research 23.6 (2022): 1-28.
13. Gan, Lige, et al. "SATTree: A SHAP-Augmented Threshold Tree for Clustering Explanation." 2023 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2023.