SHAP-C: Adapting SHAP for Interpretable Centroid-Based Clustering

Author information scrubbed for double-blind reviewing

Affiliation
email address

Abstract. As Machine Learning (ML) and Neural Networks (NNs) become more widespread across industries, there is growing interest in understanding how these systems make decisions. Despite the advantages that AI-based systems offer to industrial applications, there is increasing curiosity and concern regarding the decision-making processes of these systems. To most users, these Artificial intelligence (AI) applications remain opaque "black boxes." Trustworthy AI necessitates transparency, and Explainable Artificial Intelligence (XAI) seeks to provide this by offering interpretations based on the data used in decision-making. While many XAI models concentrate on supervised machine learning, there is a considerable need for explainability in unsupervised machine learning. Unsupervised learning methods such as clustering algorithms, by nature, uncover hidden patterns and relationships within data. However, the resulting clusters often lack context and alignment with domain expertise, hindering their practical application. This paper explores the application of SHAP (SHapley Additive exPlanations) to unsupervised learning and proposes a novel approach to enhance explainability in cluster analysis. By providing clear interpretations of cluster formations and aligning them with real-world knowledge, this method aims to foster greater trust in AI systems.

Keywords: Explainable AI (XAI) · Trustworthy AI · SHAP-C · clustering · SHAP.

1 Introduction

Explainable AI (XAI) involves enabling AI systems to deliver comprehensible explanations for the decisions and actions made. XAI seeks to bridge the gap between the advanced capabilities of AI models and the requirement for human interpretability and transparency. As AI systems become increasingly sophisticated and integrated into critical domains such as healthcare, finance, and autonomous vehicles, there is an advancing demand for AI models that offer clear and understandable explanations, which is crucial for fostering trustworthy AI.

Many AI systems are considered "black boxes" because they produce results without explaining the underlying logic. This means they make predictions or decisions without providing any insight into how these conclusions are reached. This opacity can lead to concerns about bias and discrimination in decision-making. Additionally, it hinders the adoption of AI systems in regulated industries, where clear explanations are required to comply with legal and ethical standards. This is where XAI steps in. Its aim is to provide clear and interpretable explanations accessible to both experts and non-experts. These explanations can take various forms, including textual descriptions, visualizations, or interactive interfaces. By understanding the rationale behind AI decisions, users can gain trust in the system, identify potential biases or errors, and make more informed decisions, thereby enhancing the overall trustworthiness of AI applications.

1.1 Motivation and Contribution

Traditional clustering results often lack contextual support, presenting simplistic numerical indices that can be meaningless to human interpreters. The high-dimensional nature of the data used in clustering can result in groupings that are not easily understood or implemented in real-world industrial settings. While the field of Explainable AI (XAI) has made significant progress in analyzing supervised learning tasks such as classification and regression, with approaches like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) providing instance-level explanations, these methods face challenges when applied to unsupervised learning tasks like clustering. Despite being touted as model-agnostic, these XAI techniques typically rely on labeled data as a reference point for generating explanations, which is not available in clustering scenarios. This limitation highlights the need for specialized approaches to explain clustering results in a manner that is both interpretable and applicable to real-world contexts.

Unlike supervised learning, unsupervised approaches like cluster analysis group similar data points together to identify patterns and characteristics from a machine learning perspective. Some recent research attempts to directly use the clustering assignments as the labels, hoping to leverage the existing features to interpret the clusters with these labeled data [12]. However, because the underlying models differ, the explanations generated may not be reliable for users.

To facilitate effective human-AI collaboration, increase the fidelity of the clustering assignments, and mitigate the uncertainty of the biases or errors, an explainable clustering method with adapted SHAP values is proposed in this paper. The main contributions of the study are: first, we intuitively apply SHAP to interpret centroid-based clustering approaches by defining the explanation model with unlabeled data; second, We offer local explanations for specific instances and provide global insights into the impact of input features on clusters by evaluating the proximity of instances to them; and third, we validate the value and benefits of the proposed method through the support and expertise of domain experts in enhancing industrial processes.

The remainder of the paper is organized as follows: Section 2 discusses related works on explainable clustering. Section 3 outlines the methodology, detailing the integration of the clustering and SHAP methods. In Section 4, a case study with experimental results is well analyzed. The conclusions will be summarized in the last section.

2 Related Work

2.1 Cluster Analysis

In contrast to supervised learning paradigms, such as classification which leverages labeled data for model training and subsequent explanation of predictions based on prior knowledge [3], clustering algorithms operate in the absence of such labels. the scenario of cluster analysis approaches may vary and they all aim at involving grouping similar data points together to form clusters based on their inherent similarities or relationships.

Clustering algorithms employ diverse approaches and underlying principles for forming clusters. These algorithms can be categorized into several taxonomies based on different criteria, such as centroid-based (partitioning) clustering, hierarchical clustering, and density-based clustering, etc. K-means is one of the most outstanding centroid-based algorithms, it divides the dataset into K clusters, then it iteratively assigns data points to the nearest centroid and updates the centroids based on the mean of the assigned points. Hierarchical clustering is categorized by top-down and bottom-up approaches, they are referred to as divisive and agglomeration clustering, respectively. They construct the hierarchy by iteratively merging or splitting the data points into clusters based on similarity or dissimilarity measures. Unlike centroid-based or hierarchical clustering, density-based clustering does not assume a predefined structure or a fixed number of clusters. Instead, it identifies dense regions of data points and considers them as clusters, while separating regions with low density as noise or outliers.

2.2 Explainable AI on Cluster Analysis

Interpreting clustering results within the context of trustworthy AI presents significant challenges due to the diversity of clustering criteria and the absence of ground truth data. Firstly, clustering's exploratory nature, aiming to uncover patterns without pre-existing knowledge, hinders the establishment of definitive "correctness" for cluster assignments. Secondly, the inherent ambiguity and complexity of representative encoding often lead to diverse clustering solutions depending on the chosen algorithm and its parameters. Finally, traditional clustering approaches are opaque in their reasoning. They lack built-in mechanisms to explain why data points are grouped together or how to distinguish clusters. These non-intuitive representations make it difficult to present cluster characteristics in readable explanations, thereby reducing users' trust in the model's decision-making process.

While Explainable AI (XAI) has made significant strides in supervised learning, the interpretability of unsupervised learning techniques remains an understudied domain [4]. Existing approaches ([2], [4], [6], [7], [10], [11], [12]) refer to global explanation methods, which produce an overview of the clusters to cover the majority of instances as a global explanation. providing an overview of the clusters to encompass the majority of instances. However, a small proportion of cases remain uncertain and do not fit well into these generalized rules.

[3] gives a systematic view of interpretable clustering, whereas Gilpin *et al.* attempted to differentiate interpretability and explainability in this context. Fraiman R. *et al.* [9] proposed CUBT (Clustering using unsupervised binary trees), a method utilizing binary trees for clustering. [6] constructs specified top-down decision trees to interpret clustering models by involving pre-defined features' importance as well as utilizing the splitting thresholds to guide the data partitioning process. Conversely, H. Gilpin *et al.* [5] proposed ExClus (Explainable Clustering on Low-dimensional Data Representations), which focuses on identifying influential attributes within each cluster. However, ExClus relies on hyperparameter tuning and visualizations that limit scalability and reliability. Another approach, CLTree[8], reformulates the clustering problem into a classification problem. Similarly, CLAMP (Cluster Analysis with Multidimensional Prototypes) [11] considered "supervised interpretation" for clustering explanation. It aims to depict clusters using decision rules generated from multidimensional bounding boxes representing cluster prototypes. However, both methods have limitations: CLTree interprets the decision tree path rather than the clustering model itself, and CLAMP's explanations, based on selected bounding points, can be unreliable as outliers and corner cases are not adequately addressed.

2.3 SHAP

SHAP stands for "SHapley Additive exPlanations," which is a unified framework for explaining the predictions of machine learning models, introduced by Lundberg and Lee [13]. It has gained significant popularity as a powerful tool for interpreting the black-box nature of many machine learning algorithms. SHAP is rooted in game theory and leverages Shapley values, a concept originating from cooperative game theory.

The Shapley value fairly assigns a contribution to each player in a cooperative game by considering all possible coalitions. In the context of SHAP, it assigns importance values to features based on their input values and their contribution to a model's prediction. SHAP inherits three properties from the classic Shapley values in the game theory: local accuracy, missingness, and consistency. Local accuracy requires the approximation of the explanation model to match the output of the original model as closely as possible. The missingness property describes if a feature is set to "absence" in a possible coalition, it will no longer contribute to the approximation. And the property of consistency describes if the contribution of a feature value change caused by the model to be explained, the SHAP value of that feature also changes following the same trend as the contribution change.

Specifically, in the context of explaining machine learning predictions, SHAP measures the contribution of each feature to the prediction of a particular instance by taking all possible combinations of features into account. It considers both the presence and absence of features and calculates the average marginal contribution of each feature across all those permutated coalitions. As the nature of the additive feature attribution method, the SHAP explanation model g is represented by a linear function:

$$g\left(z'\right) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{1}$$

where $z' \in \{0,1\}^M$ is the coalition vector [14], which represents the presence or absence of features ($z' = 1$ for the presence and $z' = 0$ for the abscence of feature i). M is the total number of the input features, and ϕ_i 's are the Shapley values, which quantify the importance or contribution of the corresponding features. In a coalition, $z_i' = 1$ indicates the presence of feature i, whereas the value 0 of z_i' represents the absence of feature i. The Shapley values capture the marginal contributions of each feature, considering all possible coalitions of features, and satisfy the desirable properties. It is important to note that ϕ_0 is not the Shapley value of any feature but a base value, rather a base value, defined as the expected value of the prediction when all features are absent. This additive attribution method explains how to obtain the predicted value from the base expectation by summing the contributions (Shapley values) of the present features.

When approximating the predicted values for the possible coalitions, the mapping function $h_x(z_i')$ is employed to convert the binary digits in the coalition representation to the corresponding feature values from the original feature space. Specifically, if a feature is observed ($z_i' = 1$), it is mapped to its value in the instance x. Otherwise, the recovered value is randomly sampled from the dataset's distribution for that feature. Consequently, the approximation of the explanation model can be expressed as $f(h_x(z')) = g(z')$, where f is the original machine learning model being explained, and g is the SHAP explanation model.

On the other hand, the Shapley values themselves are applied during the approximation to quantify the effect of the present features, represented by the weighted values ϕ_i. These Shapley values are computed following the principles of classic coalitional game theory:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \cdot \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right)\right] \tag{2}$$

In this equation, F represents the full set of features, and S denotes a possible coalition or subset of features from F. The Shapley value ϕ_i for feature i is calculated by summing the marginal contributions of that feature across all possible coalitions. The marginal contribution is quantified as the difference between the model's output with and without feature i, weighted by the combinatorial term $\frac{|S|!(|F|-|S|-1)!}{|F|!}$. This term ensures a fair distribution of the total contribution among all features, considering the different possible coalitions in which a feature can be present or absent. By summing these weighted marginal contributions over all coalitions, the Shapley value ϕ_i captures the overall importance or contribution of feature i to a given prediction produced by the model.

Empirically, the SHAP model typically sets the target as the probability of the predicted class in supervised learning tasks. Consequently, the explanations provided by the model aim to highlight the positive or negative influence of features by describing how they contribute to increasing or decreasing the predicted probability.

However, in unsupervised learning scenarios, there is no probabilistic function or target variable to compare against. As a result, these explanation approaches cannot be directly applied to unsupervised learning models, necessitating the development of tailored techniques for interpreting SHAP values in the unsupervised setting. Furthermore, SHAP requires a reference or baseline point to compare feature contributions against. The choice of this reference point can significantly impact the resulting explanations, as an arbitrary selection may affect the interpretability and reliability of the SHAP values.

3 Methodology

The primary goal of the proposed work is to leverage an altered SHAP approach to seamlessly provide explanations in the form of feature importance for centroid-based clustering methods. The proposed method contributes to three key aspects: First, it focuses on the selection of an appropriate centroid-based clustering algorithm, such as K-means or K-medoids, which are widely used and interpretable techniques. Second, it introduces an adjustment to the SHAP framework to enable the interpretation of clustering results, overcoming the challenges of applying SHAP to unsupervised learning scenarios. Third, it demonstrates the capability to generate both global explanations for the overall clustering structure and local explanations for individual instances within each cluster.

3.1 The selection of the clustering algorithm

However, a crucial limitation of existing interpretable models is their inability to effectively represent the results of clustering algorithms. These models typically present the clustering results solely in terms of cluster indices or labels, lacking a suitable metric to capture the underlying structure and characteristics of the clusters.

Centroid-based clustering methods, such as K-means or K-medoids, are a class of algorithms that assign data points to clusters based on the similarity of their features to the centroid of each cluster. The objective of these methods is to minimize the intra-cluster distance while maximizing the inter-cluster distance. By treating the centroid as a representative of the cluster, the distance between a data point and the cluster centroid can be utilized to estimate the similarity between the data point and the cluster, reflecting the extent of commonality between the data point and its assigned cluster.

$$f(p) = dist(C_{centroid}, p) = \sqrt{\sum_{i=1}^{M} (C_i - p_i)^2} \tag{3}$$

K-Means is one of the most widely adopted and well-established centroid-based clustering algorithms. It is a simple yet effective technique that partitions the data into K clusters based on the similarity of the data points to the cluster centroids. Without prior knowledge of the dataset, an arbitrary selection of K can introduce uncertainty and impact the subsequent processes. To fairly determine the optimal value of K, Silhouette analysis is applied in this stage for a set of possible values, which are empirically collected after a preliminary study of the dataset.

3.2 SHAP-C: Redefine SHAP for clustering interpretation

As previously discussed, SHAP specifies the explanation by approximating the prediction of the model to be interpreted. Generally, this approximation is either the probability of the predicted class in a classification task or the approximated value of the actual predicted value in a regression task. However, due to the nature of unsupervised learning and the lack of ground truth, the application of SHAP cannot directly define the objective of the explanation function.

To address this challenge and make the explanation consistent with the clustering task, the estimation of SHAP values is defined as the distance between the target data point and its corresponding cluster centroid. This distance is typically measured using a distance metric, such as Euclidean distance, Manhattan distance, or any other suitable distance function.

On the other hand, Lundberg and Lee described that the base value would be the expectation of predicted values in terms of the deterministic class in their original work [13]. In this paper, we propose using the average within-cluster sum of squares (WCSS), inspired by the Elbow method for optimizing the number of clusters in cluster analysis, as the basis for the explanation approximation.

$$f'\left(z'\right) = \frac{1}{|N|} \sum_{p_j \in N} dist\left(C_{centroid}, p_j\right) + \sum_{i=1}^{M} \phi_i z'_i \tag{4}$$

Here, $f'\left(z'\right)$ is the proposed adjusted SHAP function, N is the set of instances in cluster C, and $|N|$ is the number of instances in cluster C. $dist\left(C_{centroid}, p_j\right)$ measures the distance from the centroid to a data point p_j within the cluster. The linear weighted model $f'\left(z'\right)$ is trained by optimizing the loss function L shown as follows:

$$L\left(f, f', \pi\right) = \sum_{j \in |N|} \left[f\left(h_x^{-1}\left(z_j{}'\right)\right) - f'\left(z_j{}'\right)\right]^2 \pi\left(z_j{}'\right) \tag{5}$$

The loss function optimizes the linear model f' by utilizing the sum of squared errors with a specified coefficient π. This coefficient, also known as the SHAP kernel proposed in the original work, is used to attain the weighting that complies with the Shapley values.

3.3 Representations of the SHAP-C explanations

While SHAP has been widely introduced in various XAI literature as a local method, providing explanations for individual predictions, its properties and characteristics can be leveraged to derive global explanations as well. SHAP attributes the contribution of each feature from a base value to approximate the explainer's output. Although the contribution can be either positive or negative, indicating the promotion or demotion of the feature, the impact of the feature is reflected by its absolute Shapley value.

Additionally, adhering to the consistency property, SHAP ensures fidelity for all input features in its local explanations. This property implies that if a feature's contribution to the model's prediction changes, its Shapley value will change accordingly, reflecting the same trend. Consequently, the individual Shapley values of a feature can be utilized as a measure of its global importance across multiple instances. To determine the global importance of a feature within a cluster, the average accumulation of the absolute weighted feature values across a set of instances belonging to the cluster can be calculated:

$$I_i = \frac{1}{|N|} \sum_{j=1}^{N} |\phi_i^j| \tag{6}$$

In this equation, I_i represents the global importance of feature i within the cluster, N is the set of instances belonging to the cluster, and $|\phi_i^j|$ denotes the absolute value of the Shapley value for feature i in instance j. This global explanation, derived from the local SHAP explanations, provides insights into the most influential features that define the cluster's structure and composition, facilitating a better understanding of the clustering results.

4 Experiments

4.1 Dataset

In this section, we conduct experiments using two diverse datasets: a high-quality, publicly available UCI dataset as a benchmark [38] and a practical dataset provided by our industrial collaborator.

Seeds dataset. The first experiment utilizes the seeds dataset, provided by the Institute of Agrophysics of the Polish Academy of Sciences in Lublin [39]. This dataset comprises 210 instances, each containing seven attributes that describe the internal kernel structure of wheat. These instances are sampled from three different varieties of wheat seeds: Kama, Rosa, and Canadian. The measurements from the wheat kernels can help align agricultural and biological knowledge with the clustering process.

Industrial practical dataset. As the second experiment relies on an industrial database, a subset of data is gathered specifically related to an industrial process. Due to privacy protocols enforced by the cooperative industry providing the database, the actual names of the features are concealed and represented by standardized dummy codes. The dataset consists of 3741 records and includes seven predetermined attributes. With the guidance of experts, the appropriate number for clustering the data is verified by comparing various iterations of the clustering results; and the most influential attribute for each industrial step has been accurately identified, and the sequence of these attributes is documented as domain knowledge for the reference of the produced explanations.

4.2 Experimental Setup

K-Means, as previously discussed, serves as the primary centroid-based clustering method in this study. In the first experiment, the number of clusters is predetermined based on existing class labels. The second experiment employs a more nuanced approach, combining domain expert input with the Silhouette Coefficient metric (Rousseeuw, 1990) to determine the optimal number of clusters. The Silhouette Coefficient, applied in the second experiment, evaluates clustering quality by computing the mean Silhouette Coefficient across all data points for various cluster numbers. It takes into account both the cohesion (how close data points are to other points within the same cluster) and the separation (how far apart data points are from points in other clusters). Both determinations ensure the K-Means produces reliable assignments of the cluster for all input data. Upon completion of the clustering process, centroid information is extracted, and within-cluster distances (between data points and their respective centroids) are calculated and documented.

Following the clustering phase, the proposed SHAP-C is implemented to generate SHAP values for individual data points within each cluster. More importantly, our adapted method inherent SHAP properties, then it is capable of offering two levels of explanation: the individual level aims at revealing the contribution of features to specific cluster assignments; the global overview demonstrates feature importance across specified clusters by applying equation 6 to calculate average accumulated SHAP values for each feature.

4.3 Experimental Results

Table 1. Comparative performance metrics for clustering algorithm on the seeds dataset

Cluster Number	Size	Precision	Recall	F1
1	70	0.85	0.81	0.83
2	70	0.98	0.86	0.92
3	70	0.85	1.00	0.92

As described in the UCI repository, the three numerical labels in the Seeds dataset indicate Kama, Rosa, and Canadian wheat varieties, respectively. Table 1 compares the performance of the clustering model across these diverse wheat varieties. Cluster 2 has the highest precision (0.98), indicating that 98% of the instances assigned to this cluster are likely from the same wheat variety. Cluster 3 captures all instances of one variety (perfect recall) but includes some instances from other varieties (lower precision). Cluster 1 seems to be the most challenging to classify accurately, possibly representing the variety with features that overlap more with the others.

Different from the feature analysis based on conventional statistical analyses, the adapted SHAP values serve as indicators of each feature's contribution to the clustering process. Table 2 presents an analysis of feature contributions for each cluster in the Seeds dataset using the mean absolute SHAP values. The features are ranked according to their average SHAP values, indicating their impact on the clustering process for each corresponding wheat variety.

The feature rankings show some variation across the three clusters, but notably, $Asymmetry\ coefficient$ and $Area$ consistently emerge as the top two contributing features in all clusters. This suggests that the model heavily relies on these two characteristics when making clustering decisions. Unlike Cluster 2 and Cluster 3, Cluster 1 shows a more even distribution of importance among its top four influential features, with only minor differences in their mean SHAP values. This lack of highly distinctive features might explain the slightly lower performance observed for Cluster 1 in previous analyses. Additionally, Contrary to expectations based on previous research [40], which implied that kernel-related features (width, length, and groove length) are crucial for wheat variety identification, our experimental results show a different pattern. These kernel-related features generally show lower average SHAP values, indicating less contribution to the clustering process than anticipated.

To accurately distinguish how the selected features contribute to the clustering process, we need to look beyond individual instances and consider the overall attribute contributions. The experimental results referred to Table3. The attributes in each subsection are also sorted based on the average SHAP absolute values, so the rank reflects the average contribution of the attributes when doing clustering. In this study, we benefit from collaboration with industrial domain experts. Their feedback on our clustering analysis allows us to compare decision-making strategies between human expertise and AI-driven insights, providing a valuable perspective on the practical implications of our results.

The first section of the table provides a global overview of attribute contributions across all records in the dataset. $Feature\_4$, $feature\_7$, and $feature\_6$ emerge as the most influential in the clustering process. According to our domain experts, these three attributes are associated with the final stages of the production process. This insight indicates that the finishing stages play a significant role in product manufacturing throughout the entire industrial workflow.

Feature Contributions in Cluster 1		
Ranking	Feature	mean_abs_shap
1	Asymmetry_coefficient	0.273257
2	Area	0.123926
3	Width_of_kernel	0.081337
4	Length_of_kernel_groove	0.073772
5	Length_of_kernel	0.049213
6	Perimeter	0.043966
7	Compactness	0.014615
Feature Contributions in Cluster 2		
Ranking	Feature	mean_abs_shap
1	Asymmetry_coefficient	0.506818
2	Area	0.124254
3	Width_of_kernel	0.05329
4	Length_of_kernel_groove	0.04211
5	Length_of_kernel	0.040932
6	Perimeter	0.034227
7	Compactness	0.008723
Feature Contributions in Cluster 3		
Ranking	Feature	mean_abs_shap
1	Asymmetry_coefficient	0.485764
2	Area	0.089792
3	Width_of_kernel	0.073262
4	Perimeter	0.029747
5	Compactness	0.027634
6	Length_of_kernel_groove	0.026165
7	Length_of_kernel	0.022423

Table 2. Features contributions in each cluster.

Global Attribute Contributions		
Ranking	**Feature**	**mean_abs_shap**
1	feature_4	0.15709
2	feature_7	0.09587
3	feature_6	0.08819
4	feature_2	0.07866
5	feature_1	0.07473
6	feature_3	0.03877
7	feature_5	0.00227
Attribute Contributions of Cluster 3		
Ranking	**Feature**	**mean_abs_shap**
1	feature_4	0.10812
2	feature_7	0.09747
3	feature_6	0.08335
4	feature_2	0.05572
5	feature_1	0.05020
6	feature_3	0.05013
7	feature_5	0.00377
Attribute Contributions of Cluster 7		
Ranking	**Feature**	**mean_abs_shap**
1	feature_4	0.26592
2	feature_2	0.08837
3	feature_1	0.08674
4	feature_7	0.07895
5	feature_6	0.07822
6	feature_3	0.06239
7	feature_5	0.00197

Table 3. The attribute contributions of the groups are calculated by averaging the absolute SHAP values for each feature within each group.

On the other hand, The two rest subtables in Table3 exhibit the rankings of the feature contribution on Cluster 3 and Cluster 7, offering insights into the specific characteristics of these product groups.

For Cluster 3, while the overall ranking of features aligns closely with the global explanation, the specific SHAP values differ, reflecting the unique attributes of this cluster. The similarity in feature ranking to the global overview suggests that Cluster 3 likely represents products that closely follow the typical production process. These could be considered the "regular" products in the manufacturing line.

Cluster 7, shown in the bottom subtable, presents a notably different feature distribution. While $feature\_4$ remains the most important, $feature\_2$ and $feature\_1$ take the second and third places, respectively. Our domain experts have identified these features as being related to the early stages of the industrial production process. This shift in feature importance suggests that for products in Cluster 7, the initial stages of production have a more significant impact on the final outcome than in other clusters.

The distinct feature ranking in Cluster 7 implies that the characteristics of products in this category may be influenced by additional factors not fully captured in the clustering process. This unique profile suggests that special attention should be given to the initial stages of the production process when manufacturing products that fall into this category.

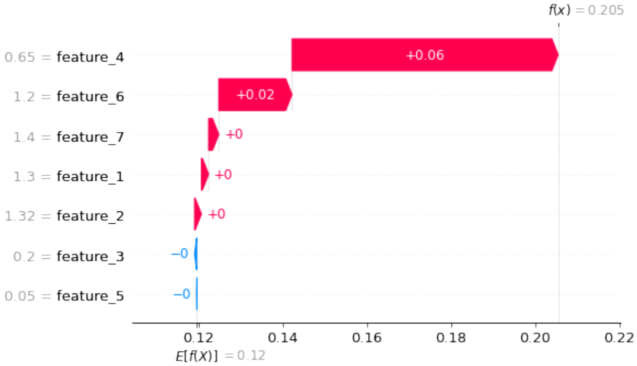

Fig. 1. The SHAP values for an individual explanation on clustering a specified instance into the group. The baseline $E[f(x)]$ shows the average within-cluster distance of the cluster. The figure presents how the feature contribution is distributed when getting to the nearest cluster centroid from the current position.

Applying our proposed method to the practical dataset related to the industrial process allows us to gain insights into individual instances. Figures 1 and 2 illustrate SHAP-C individual explanations for two different records, showcasing the contribution of all input features during the clustering process.

Figure 1 represents an instance assigned to Cluster 0 with a minimal intra-cluster distance of 0.205 across all possible clusters. The diagram uses colored directional bars to illustrate how each attribute contributes to positioning the instance relative to the cluster centroid, starting from the base value $E[f(x)]$. Red bars indicate features that increase commonality with the cluster, while blue bars represent features that differentiate the instance from the cluster's typical characteristics. In this case, $feature\_4$ stands out with a large positive SHAP value, strongly contributing to the record's membership in the cluster.

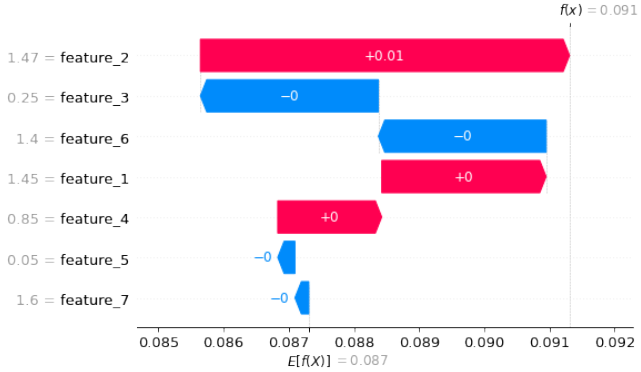

Fig. 2. Another individual explanation of an instance in Cluster 3.

Conversely, $feature\_3$ and $feature\_5$, represented by short blue bars, slightly push the instance away from the cluster centroid.

Figure 2 presents a different instance from Cluster 3, which has drawn particular attention from domain experts due to its unexpected assignment to this cluster. This case study demonstrates the diverse distribution of attribute contributions and highlights the value of combining machine learning insights with expert knowledge. Upon closer examination, the small difference of 0.004 between the base value and the within-cluster distance indicates that this record is very close to the cluster centroid, despite the experts' initial surprise at its categorization. This proximity suggests that the clustering algorithm has identified similarities that may not be immediately apparent through traditional analysis.

In the figure, $Feature\_2$ emerges as the most influential attribute, strongly driving the record toward the cluster center. This insight could provide valuable information about a key characteristic that defines this cluster, potentially revealing new patterns or relationships in the production process. Notably, more than half of the features have a negative impact on the assignment process, slightly pulling the instance away from the centroid. This complex interplay of features pushing and pulling the instance within the cluster space illustrates the nuanced nature of the clustering process and the multifaceted characteristics of the products.

Comparing both diagrams reveals that even within the same cluster, the contribution of attributes can vary significantly between instances. This variability highlights the complexity of the clustering process and the subtle differences between products within the same category. Unlike global explanations, these individual SHAP explanations offer a machine learning perspective that allows for a detailed understanding of why specific products are assigned to certain categories. They provide insights that complement and sometimes challenge traditional domain knowledge, offering a fresh view of product characteristics and their relationships.

Specifically, these instance-level explanations assist in identifying potential outliers or unique products within a cluster, which might not be apparent through conventional analysis. The experts may be able to take care of the instances where machine learning insights differ from traditional domain expertise, prompting further investigation.

5 Conclusion

Machine learning models are extensively employed in real-world applications, yet their outputs can often be confusing or incomprehensible, even to domain experts, despite satisfactory performance metrics. This research investigates the challenges of explaining unsupervised learning models, particularly clustering algorithms. The issue that renders most existing model-agnostic explanation methods unavailable for tasks like clustering is addressed and discussed. Due to the lack of prior knowledge, clustering models may provide inaccurate estimations, and their results are difficult to interpret and validate.

In this study, we enhance transparency in AI by proposing an adapted method SHAP-C for explaining centroid-based clustering. This approach aims to achieve three key objectives: first, it extends the flexibility of model-agnostic methods to effectively explain clustering results; second, it elucidates feature attribution in the cluster assignment process, providing clarity on the decision-making mechanisms; and third, it aggregates individual explanations to summarize the influential features that formed clusters based on their contributions. By addressing these aspects, our method promotes trust and transparency in AI systems, ensuring that clustering algorithms are more comprehensible and reliable for users.

Subsequently, the proposed method is implemented on a real-world industrial dataset, and the resulting explanations offer an intuitive perspective on analyzing clustering results with the assistance of domain expertise. Overall, SHAP-C offers a more transparent view of the clustering model, making the resulting clusters understandable and meaningful to users, rather than requiring manual interpretation of individual records to conclude cluster properties.

Considering the experimental conditions, future research will focus on the following topics. First, as the explainability of the model may not be quantified, there is a need to establish a metric to evaluate the performance of explanation models. Second, current explanation approaches produce results from geometric or statistical perspectives. Further investigation will focus on involving additional knowledge bases to improve the model's explainability and align it with domain knowledge and real-world contexts.

References

1. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. " Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

2. Dasgupta, Sanjoy, et al. "Explainable k-means and k-medians clustering." arXiv preprint arXiv:2002.12538 (2020).
3. Yang, Haoyu, Lianmeng Jiao, and Quan Pan. "A survey on interpretable clustering." 2021 40th Chinese Control Conference (CCC). IEEE, 2021.
4. Ellis, Charles A., et al. "Algorithm-agnostic explainability for unsupervised clustering." arXiv preprint arXiv:2105.08053 (2021).
5. L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA), Turin, Italy, Oct. 2018, pp. 80–89.
6. Gan, Lige, et al. "SATTree: A SHAP-Augmented Threshold Tree for Clustering Explanation." 2023 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2023.
7. Vankwikelberge, Xander, et al. "ExClus: Explainable Clustering on Low-dimensional Data Representations." arXiv preprint arXiv:2111.03168 (2021).
8. Liu, Bing, Yiyuan Xia, and Philip S. Yu. "Clustering via decision tree construction." Studies in Fuzziness and Soft Computing 180 (2005): 99.
9. Fraiman, Ricardo, Badih Ghattas, and Marcela Svarc. "Clustering using Unsupervised Binary Trees: CUBT." arXiv preprint arXiv:1011.2624 (2010).
10. Kuk, Michał, Szymon Bobek, and Grzegorz J. Nalepa. "Explainable clustering with multidimensional bounding boxes." 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2021.
11. Bobek, Szymon, et al. "Enhancing cluster analysis with explainable AI and multidimensional cluster prototypes." IEEE Access 10 (2022): 101556-101574.
12. Saisubramanian, Sandhya, Sainyam Galhotra, and Shlomo Zilberstein. "Balancing the tradeoff between clustering value and interpretability." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.
13. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
14. Molnar, Christoph. Interpretable machine learning. Lulu. com, 2023.
15. Islam, Sheikh Rabiul, et al. "Explainable artificial intelligence approaches: A survey." arXiv preprint arXiv:2101.09429 (2021).
16. Thorndike, Robert. "Who belongs in the family?." Psychometrika 18.4 (1953): 267-276.
17. Reddy, Y. Subba, and P. Govindarajulu. "An efficient user centric clustering approach for product recommendation based on majority voting: a case study on wine data set." IJCSNS 17.10 (2017): 103.
18. Cao, Bin, et al. "Domain knowledge-guided interpretive machine learning: formula discovery for the oxidation behavior of ferritic-martensitic steels in supercritical water." Journal of Materials Informatics 2.2 (2022): 4.
19. Liu, Bing, Yiyuan Xia, and Philip S. Yu. "Clustering via decision tree construction." Studies in Fuzziness and Soft Computing 180 (2005): 99.
20. Meng, Yuan, et al. "What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values." Journal of Theoretical and Applied Electronic Commerce Research 16.3 (2020): 466-490.
21. Futagami, Katsuya, et al. "Pairwise acquisition prediction with SHAP value interpretation." The Journal of Finance and Data Science 7 (2021): 22-44.
22. Li, Ziqi. "Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost." Computers, Environment and Urban Systems 96 (2022): 101845.

23. Ekanayake, I. U., D. P. P. Meddage, and Upaka Rathnayake. "A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP)." Case Studies in Construction Materials 16 (2022): e01059.
24. Mokhtari, Karim El, Ben Peachey Higdon, and Ayşe Başar. "Interpreting financial time series with SHAP values." Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering. 2019.
25. Gelbard, Roy, Orit Goldman, and Israel Spiegler. "Investigating diversity of clustering methods: An empirical comparison." Data & Knowledge Engineering 63.1 (2007): 155-166.
26. Baptista, Marcia L., Kai Goebel, and Elsa MP Henriques. "Relation between prognostics predictor evaluation metrics and local interpretability SHAP values." Artificial Intelligence 306 (2022): 103667.
27. Loecher, Markus. "Debiasing MDI Feature Importance and SHAP Values in Tree Ensembles." Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, August 23–26, 2022, Proceedings. Cham: Springer International Publishing, 2022.
28. Shen, Yang, et al. "An automatic visible explainer of geometric knowledge for aeroshape design optimization based on SHAP." Aerospace Science and Technology 131 (2022): 107993.
29. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
30. Weerts H J P, van Ipenburg W, Pechenizkiy M. A human-grounded evaluation of sḫap for alert processing[J]. arXiv preprint arXiv:1907.03324, 2019.
31. Šarčević, Ana, et al. "Cybersecurity Knowledge Extraction Using XAI." Applied Sciences 12.17 (2022): 8669.
32. Dickinson, Quinn, and Jesse G. Meyer. "Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences." PLOS Computational Biology 18.1 (2022): e1009736.
33. Wang, Dong, et al. "Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods." Journal of Environmental Management 301 (2022): 113941.
34. Bowen, Dillon, and Lyle Ungar. "Generalized SHAP: Generating multiple types of explanations in machine learning." arXiv preprint arXiv:2006.07155 (2020).
35. Li, Richard, et al. "Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival." JCO Clinical Cancer Informatics 4 (2020): 637-646.
36. Pereira, Filipe Dwan, et al. "Explaining individual and collective programming students' behavior by interpreting a black-box predictive model." IEEE Access 9 (2021): 117097-117119.
37. Wang Y, Mase M, Egi M. Attribution-based Salience Method towards Interpretable Reinforcement Learning[C]//AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1). 2020.
38. DheeruDuaandCaseyGraff.2017.UCIMachineLearningRepository. http://archive.ics.uci.edu/ml
39. The Seeds Dataset. https://archive.ics.uci.edu/dataset/236/seeds
40. Charytanowicz, Małgorzata, et al. "Complete gradient clustering algorithm for features analysis of x-ray images." Information Technologies in Biomedicine: Volume 2. Springer Berlin Heidelberg, 2010.