Explaining Clusters Using Minimal Weighted Edge Coverage

Abstract

Current clustering techniques in unsupervised learning lack interpretability. This is primarily because clustering represents a combinatorial optimization problem that becomes exponentially complex in large-dimensional spaces. Consequently, most clustering algorithms employ intricate mathematical computations, statistical assumptions, distance approximations, and data transformations in ways that diminish their interpretability. This study introduces a Linear Integer Programming model to optimize the balance between interpretability and quality, drawing inspiration from the graph theory's Minimal Edge Covering problem. An edge cover is a set of graph edges ensuring each vertex of the graph is incident to at least one edge of the set; the challenge lies in determining the smallest set possible. By adopting this approach, data can be grouped into clusters in the form of tree-like structures, enhancing our comprehension of the clustering process. If the edges are weighted to represent dissimilarities or distances, the problem becomes the Minimum Weighted Edge Covering (MWEC) problem.

1 Introduction

Interpretability is increasingly essential in AI systems, particularly in high-risk areas, as it ensures outcomes are reliable for strategic and critical decisions. This necessity is crucial in clustering analysis because of its unsupervised nature. For example, medical experts are often skeptical of data-driven models due to the lack of their explainability [1]. Clustering is an unsupervised learning method used across various fields to identify heterogeneous sub-populations within a sample. The interpretability of clustering methods can be challenging for several reasons:

1. *Lack of Ground Truth*: Clustering is unsupervised, meaning there are no predefined labels or categories to guide the algorithm. This lack of ground truth makes it hard to validate and interpret the clusters.
2. *Complexity of Algorithms*: clustering represents a combinatorial optimization problem that becomes exponentially complex in large-dimensional spaces. Consequently, most clustering algorithms employ intricate mathematical computations, distance approximations, and data transformations in ways that diminish their interpretability.
3. *High-Dimensional Feature Space*: Clustering often occurs in high-dimensional spaces, where understanding the relationships between features and clusters can be challenging. High-dimensional data and dimension reduction techniques can obscure the meaning of clusters.
4. *Distance Metrics*: Clustering relies on distance metrics to group similar data points. The choice of metric can significantly impact the clustering results, and understanding why certain points are grouped together based on these metrics is not always straightforward.

5. *Cluster Shape and Size:* Real-world data can produce clusters of varying shapes and sizes, which may not align with human intuition. For example, some algorithms assume spherical clusters, which may not be suitable for all datasets.
6. *Overlapping Clusters:* Clusters can overlap or have ambiguous boundaries, making it difficult to interpret clear separations between them.
7. *Algorithm-Specific Parameters:* Many clustering methods require setting parameters (e.g., the number of clusters in k-means). The selection of these parameters can affect the results, and interpreting why certain parameter choices work better than others can be non-trivial.
8. *Lack of Contextual Information:* Clusters are formed based on the data features alone, without considering external or contextual information that might provide a clearer understanding of the clusters.

This research aims to address reasons #2, #5 and #6 among the aforementioned factors.

1.1 Literature Review

Present approaches to interpreting or explaining clustering rely extensively on statistical inference, distributional assumptions, hybris models, or post-modeling agnostic tools. Such a statistical perspective can make it difficult, if not impossible, to comprehend why a specific data point is assigned to a particular cluster? Likewise, it would be difficult to answer counterfactual questions like what if the distance between given data point with its neighbors change a bit? Another challenge with the statistical approaches is the need for implementing additional models and extra assumptions, requiring added layer of explainability to present the results in a manner understandable to humans. Besides, most explainable clustering techniques are focused on centroid-based algorithms which works well when clusters are linearly separable, compact, and spherical shape. For example, Moshkovitz et al. (2020) stated that, measuring cluster quality by the k-means and k-medians objectives, there must exist a tree-induced clustering whose cost is comparable to that of the best unconstrained clustering [2]. They defined the price of explainability for a clustering task as the unavoidable loss, in terms of the objective function, if we force the final partition to be explainable. They proposed a threshold tree approach where an explainable clustering is given by a partition, induced by the leaves of a decision tree, that optimizes k-means objective function. To doing so, the constructed centroid-based clusters must be linearly separable to be explained by a decision tree. Laber and Murtinho (2021) extended the above framework for k-centers and maximum-spacing problems [3].

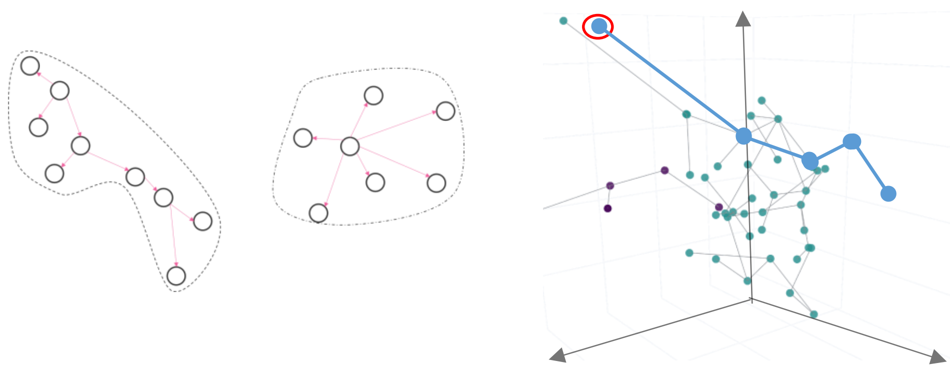
Among the highly esteemed hybrid statistical methods and post-modeling tools, Spotify Engineering team developed an explainable Clustering method: Recursive Embedding and Clustering [4]. In this method first the low-dimensional representation of the original data is constructed using UMAP (Uniform Manifold Approximation and Projection) and then clusters are created using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Finally, an XGBoost is trained using the raw data as input and labels (HDBSCAN classified as output) and understand the feature contribution using SHAP values. Likewise, Shan (2023) studied approximation algorithms for explainable k-medians and k-means clustering. The goal was to find a threshold decision tree that partitions data into k clusters and minimizes the k-medians or k-means objective. The obtained clustering is easy to interpret because every decision vertex of a threshold tree splits the vertex into two groups with a threshold cut on a single feature. The price of explainability is defined as the ratio of its cost and the optimal unconstrained cost [5]. Prabhakaran et. al. (2022) proposed Explainable K-means clustering (ExKMC) algorithm for occupancy estimation. ExKMC by default creates a small tree with k leaves that partitions the data into k clusters, and it also outputs a new tree with k' leaves where k'≥ k that provides explainable clusters. This method makes a simple trade-off between the accuracy of prediction and the interpretability of the clustering decisions [6]. Deshmak et al. (2023) proposed an improved hybrid classical-quantum clustering (qk-means – running k-means on a quantum computer) Model. This model uses learning strategies such as the Local Interpretable Model-agnostic Explanations (LIME) method and improved qk-means algorithm to diagnose abnormal activities based on breast cancer images and Knee Magnetic Resonance Imaging (MRI) datasets to generate an explanation of the predictions [1]. Turfah and Wen (2024) introduced a Distinguishability criterion, measuring the overall separability of a given cluster configuration. This criterion is derived by quantifying the misclassification probability from a multi-class classification problem. This criterion is naturally interpreted as the probability of misclassifying a data point under the given cluster configuration [7]. Similarly, Alvarez-Garcia et al. (2024) used a classification model in combination with a clustering method to enhance explainability and classify future data points. The labels generated during the classification phase will subsequently be utilized for interpretability via Shapley values [8]. Guilbert et al, (2024) proposed a framework in which an explanation of a cluster is a set of patterns (a set of descriptors). They proposed a constrained clustering method for declarative clustering with Explainabilty-driven Cluster Selection (ECS) that integrates structural or domain expert knowledge expressed by means of constraints. The key idea is that a good global explanation of a clustering should give the characteristics of each cluster taking into account their abilities to describe its objects (coverage) while distinguishing it from the other clusters (discrimination). Their method heavily relies on expert knowledge and provided descriptors [9]. Chen and Güttel (2024) introduced a clustering technique known as CLASSIX, which provide textual explanation why two data points belong to the same cluster or why they are in separate clusters. However, this claim is not clearly substantiated in the main text of the article [10].

Several research studies focus on explainable-by-design clustering, where the structure of the clusters inherently provides interpretability and explainability. For instance, Davidson et. al. (2022) proposed an clustering approach that not only finds clusters but also exemplars to explain each cluster. They say that an instance x explains another instance y (or instance x serves as an exemplar for instance y) if y falls within the ball of radius ε centered at x. Exemplars are a natural mechanism for explanation of concepts by enumerating the different variations of the concept. Their setting was naturally a bi-objective clustering problem with respect to cluster quality and explanation quality [11].

1.2 Contribution

the intuition behind the proposed model is relatively straightforward since interpretability is inherently present within the cluster structure, eliminating the need for extra models, assumptions, and tools. Inspired by Davidson et. al. (2022), when provided with a desired number of clusters, the proposed model organizes data points into clusters with tree (skeleton)-like structures. The structure of each cluster represents Minimum Spanning Tree (MST) of the cluster where each parent vertex acts as an exemplar for its children. The idea is that a child vertex differs only slightly from its exemplar in the feature space. The tree-like structure guarantees a unique parent and thereby a unique exemplar chain for each vertex as illustrated in Figure 1. The exemplar chain is a series of distinct ancestors that result in a particular vertex being part of the cluster. The proposed model does not constrain the length of exemplar chains, allowing the clusters to take more irregular, non-convex envelops, as illustrated in Figure 1. Here are other advantages of our model:

- It can justify potential overlapping clusters with ambiguous boundaries or non-convex forms.
- Lack of sensitivity to the outliers.
- The leaves to parents (LP) ratio helps us understand the connectivity among points in a cluster and thereby cluster's geometric shape. A high LP ratio signifies a hub-like cluster with one parent explaining all points, while a low LP ratio indicates a long exemplar chain. The direction of exemplar chain within the feature space shows which features and to what extent explain the vertices along the chain, as shown in Figure 2. Determining this direction is beyond the scope of this research and is suggested for future investigation.



Clusters with various geometric shapes — Exemplar chain of a given point

Figure 1: Clusters with tree-like structure

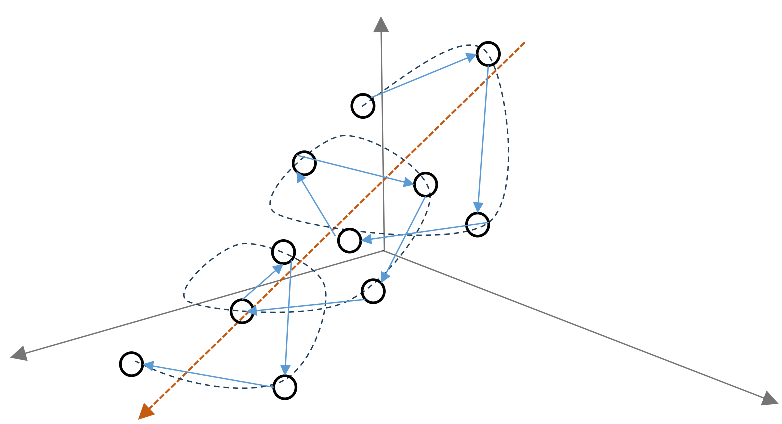


Figure 2: A cluster in the form of a long-directed exemplar chain in feature space

The proposed model is inspired by Minimal Weighted Edge Covering problem in graph theory. An edge cover is a set of graph edges ensuring each vertex of the graph is incident to at least one edge of the set. A minimum edge covering is an edge covering of smallest possible size. If the edges are weighted (e.g., dissimilarities/distances), the problem becomes the Minimum Weighted Edge Covering (MWEC) problem.

MWEC differs from the Minimum Spanning Tree (MST) problem. MST problem aims to find a subset of edges in a connected, edge-weighted undirected graph that links all vertices without cycles and with minimal total edge weight. While MST problem connects all the points in one giant tree-like structure, MWEC problem can produce multiple such clusters each of which represents an MST. Alternatively, MWEC can be constructed by cutting longer edges in the MST to split the points into separate clusters. When the number of clusters (K) equals one, the proposed model reduces to MST problem.

The proposed model provides an exact solution via Linear Binary Programming (LBP) which is more tractable compared to the exact solution generated by conventional Quadratic Binary Programming [12].

2 Proposed Mathematical Model

The minimal edge covering problem involves grouping edges rather than clustering vertices. To construct the model, the symmetric weight/distance matrix of data points can be reorganized into an 1D array of size $L = \frac{N(N-1)}{2}$ where N is the number of data points and index $l = i\left(N - \frac{i+1}{2}\right) - N + j$ is equivalent to entity (i, j) in the weight matrix, as shown in Figure 2.

	j =1	j =2	j =3	j =4
i =1				
i =2	$l = 1$			
i =3	$l = 2$	$l = 4$		
i =4	$l = 3$	$l = 5$	$l = 6$	

Figure 2: Mapping symmetric weight matrix into 1D array

Each edge l is labeled by weight d_l, origin vertex $O(l) = i$ and destination vertex $D(l) = j$. Having the above notations, the proposed LBP model can be formulated as follows:

$$\min Z = \sum_{l=1}^{L} x_l d_l, \tag{1}$$

s.t:

$$\sum_{O(l)=i} x_l + \sum_{D(l)=i} x_l \geq 1 \quad \forall i \tag{2}$$

$$\sum_{D(l)=i} x_l \leq 1 \quad \forall i, \tag{3}$$

$$\sum_{l=1}^{L} x_l = K\left(\left\lfloor \frac{N}{K} \right\rfloor - 1\right) + MOD(N, K), \tag{4}$$

$$x_l \in \{0,1\}, \tag{5}$$

Where binary variable $x_l = 1$ means edge l belongs to MWEC; otherwise, $x_l = 0$. Objective (1) calculates the total cost of constructing MWEC in terms of weighted edges. Constraints set (2) ensure that all data points are covered. Constraint set (3) ensures that each data point has a unique parent and avoids cycles in the MWEC. Equity (4) is sparsity constraint to control the number of clusters formed. It can be demonstrated without much difficulty that the number of edges in MWEC given K clusters is $K\left(\left\lfloor\frac{N}{K}\right\rfloor - 1\right) + MOD(N, K)$. Constraint (5) addresses the binary variables integrality.

Objective function (1) isn't a standard clustering objective aimed at minimizing within-cluster distances or maximizing inter-cluster discrimination which are more suitable for linearly separable centroid-based clusters. Instead, it is to construct MWEC where the number of edge groups (clusters) is already known. Therefore, the proposed model is not suitable to determine the optimal number of clusters.

Model (1-5) is NP-hard with approximate complexity $O(2^M)$ where $M = \frac{N(N-1)}{2}$ is the number of binary variables. The number of constraints, i.e., $2N + 1$, will affect the complexity depending on the algorithm used. The model can be solved using the classical exact algorithms such as Branch-and-Bound, Branch and Cut or cutting planes.

3 Experimental Results

The performance of the proposed model is compared to the k-means and spectral clustering methods from scikit-learn package via two publicly available datasets: 1- Client Credit Card Activity with five numerical features (Figure 3), and 2- Customer Segmentation based on demographic information given seven features with mixed datatypes (binary, categorical and numerical), as shown in Figure 4. While distance d_l in the first dataset is calculated using Euclidean metric, it is calculated using GOWER [13] metric in the second dataset. GOWER uses "Manhattan" distance for continuous variables and "dice" distance for measuring similarity between non-continuous variables. Spectral clustering is preferred because it effectively handles clusters with potentially non-convex structures. Since k-means does not support GOWER metric, only spectral clusters will be provided for dataset with mixed datatypes. The quality of constructed clusters is measured in terms of Silhouette metric. The silhouette value is a measure of how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The proposed mathematical model is solved using CBC (COIN Branch and Cut) algorithm – an open-source mixed-integer programming solver embedded in PULP python package [14].

Table 3 shows the Silhouette values for the MWEC model compared to K-means and Spectral clusters for 21 random samples from two datasets. The table evidences that the clustering quality of our model is highly competitive against the other two unexplainable methods; especially given the samples with mixed datatypes (six bottom samples). Figure 3 shows three clusters with tree-like structure generated by MWEC model in a feature space reduced by PCA method for the sack of visualization. As is evident in this figure, there are two potential overlapping clusters. In this scenario, the MSP of each cluster can be used to explain the ambiguous boundaries between the two clusters.

Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
100000	2	1	1	0
50000	3	0	10	9
50000	7	1	3	4
30000	5	1	1	4
100000	6	0	12	3

Table 1: Sample data - Client Credit Card Activity (Non-mixed Data types) – Euclidian Distance

ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
100000001	0	0	67	2	124670	1	2
100000002	1	1	22	1	150773	1	2
100000003	0	0	49	1	89210	0	0
100000004	0	0	45	1	171565	1	1
100000005	0	0	53	1	149031	1	1

Table 2. Sample Data - Customer Segmentation based on Demo. Info. (Mixed Datatype) – Gower Distance

N	*L*	*K*	Mixed Datatype	*K*-means Silhouette	Spectral Silhouette	MWEC Silhouette	MWEC - $\sum_{l=1}^{L} d_l$	MWEC-CPU (min.)
66	2145	2	FALSE	0.400	0.402	**0.527**	64	0.079
66	2145	3	FALSE	**0.530**	0.334	0.469	63	0.077
66	2145	4	FALSE	**0.412**	0.341	0.361	62	0.072
132	8646	2	FALSE	0.425	0.425	**0.502**	130	0.553
132	8646	3	FALSE	**0.508**	0.473	0.469	129	0.561
132	8646	4	FALSE	**0.412**	0.379	0.356	128	0.604
198	19503	2	FALSE	0.432	0.176	**0.492**	196	1.866
198	19503	3	FALSE	0.520	**0.524**	0.477	195	1.841
198	19503	4	FALSE	**0.403**	**0.403**	0.341	194	2.029
264	34716	2	FALSE	0.434	**0.494**	**0.494**	262	4.812
264	34716	3	FALSE	**0.521**	**0.521**	0.457	261	4.791
264	34716	4	FALSE	**0.410**	0.406	0.328	260	4.808
264	34716	5	FALSE	**0.331**	0.322	0.296	259	4.901
330	54285	2	FALSE	0.429	**0.485**	**0.485**	328	9.299
330	54285	3	FALSE	**0.522**	**0.522**	0.456	327	9.38
200	19900	2	TRUE	NA	-0.007	**0.212**	198	1.955
200	19900	3	TRUE	NA	-0.027	**0.229**	197	1.931
200	19900	4	TRUE	NA	-0.049	**0.141**	196	1.932
400	79800	2	TRUE	NA	-0.072	**0.185**	398	15.052
400	79800	3	TRUE	NA	-0.200	**0.202**	397	14.939
400	79800	4	TRUE	NA	-0.037	**0.153**	396	15.585

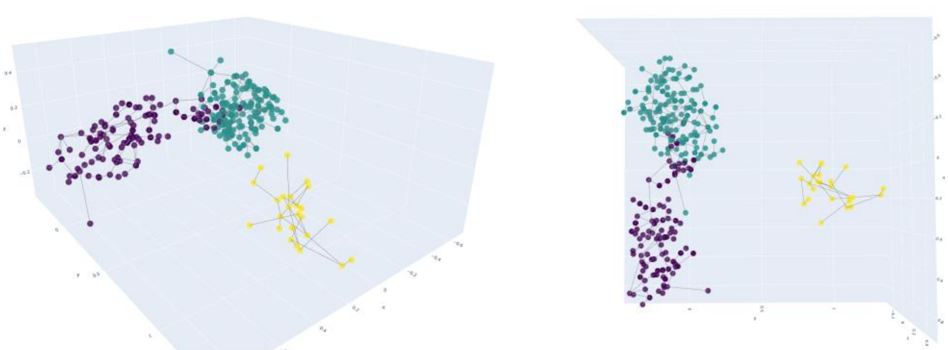


Figure 3: Clusters created by MWEC model with tree-like structure/connectivity

4 Conclusions

This research indicates that the connectivity between points within a cluster, represented by the minimal spanning tree (MSP), is fundamentally explainable without need to the extra post-clustering models or tools. This concept is backed by MSP's parent-child relationship, where each parent vertex serves as an exemplar for its children, implying that a child vertex varies only slightly from its parent (exemplar) in the feature space. This idea has several advantages: It can justify potential overlapping clusters with ambiguous boundaries, lack of sensitivity to the outliers, and leaves-to-parents ratio of MSP helps us understand the connectivity among points in a cluster and thereby cluster's geometric shape. The experimental results indicate that the clustering quality of the proposed approach is highly competitive against conventional unexplainable clustering methods, especially given the samples with mixed datatypes.

References

[1] Deshmukh, S., Behera, B.K., Mulay, P., Ahmed, E.A., Al-Kuwari, S., Tiwari, P., Farouk, A (2023) Explainable quantum clustering method to model medical data, Knowledge-Based Systems, 267: 110413.

[2] Moshkovitz, M., Dasgupta, S., Rashtchian, C., Frost, N (2020) Explainable k-means and k-medians clustering, in Proceedings of the 37th International Conference on Machine Learning, H. D. III and A. Singh, eds., vol. 119 of Proceedings of Machine Learning Research, PMLR, 13–18: 7055–7065.

[3] Laber, E. and Murtinho, L. (2021) On the price of explainability for some clustering problems. International Conference on Machine Learning. 5915-5925 PMLR.

[4] Pereira, G. Recursive Embedding and Clustering, Spotify R&D and Engineering, December 2023.

[5] Shan, L. Approximation Algorithms for Explainable Clustering, PhD Dissertation, Northwestern University, Evanston, Illinois, September 2023.

[6] Prabhakaran, K., Dridi, J., Amayri, M., Bouguila, N. (2022) Explainable K-Means Clustering for Occupancy Estimation, Procedia Computer Science, 203: 326-333.

[7] Turfah, A., Wen, X. (2024) Interpretable Clustering with the Distinguishability Criterion, https://arxiv.org/abs/2404.15967v2.

[8] Alvarez-Garcia, M., Ibar-Alonso, R., Arenas-Parra, M. (2024) A comprehensive framework for explainable cluster analysis, Information Sciences, 663.

[9] Guilbert, M., Vrain, C., Dao, TBH (2024) Towards Explainable Clustering: A Constrained Declarative based Approach, https://arxiv.org/abs/2403.18101.

[10] Chen, X., Güttel, S (2024) Fast and explainable clustering based on sorting, Pattern Recognition, Volume 150, https://doi.org/10.1016/j.patcog.2024.110298

[11] Davidson, I., Livanos, M., Gourru, A., Walker, P., Velcin, J., Ravi, S.S. (2022) Explainable Clustering via Exemplars: Complexity and Efficient Approximation Algorithms, https://arxiv.org/abs/2209.09670.

[12] Bonizzoni, P., Vedova, G.D., Dondi, R., Jiang, T (2008) On the approximation of correlation clustering and consensus clustering, J. Comput. Syst. Sci., 74(5): 671-696.

[13] Gower, Jhon C. (1971) A general coefficient of similarity and some of its properties, Biometrics. 27(4): 857–871.

[14] The Computational Infrastructure for Operations Research (COIN) project, Stanford University, link: https://github.com/coin-or/Cbc.