Interpretable Clustering via Optimal Classification Trees

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

State-of-the-art clustering algorithms use heuristics to partition the feature space and provide little insight into the rationale for cluster membership, limiting their interpretability. In healthcare applications, the latter poses a barrier to the adoption of these methods since medical researchers are required to provide detailed explanations of their decisions in order to gain patient trust and limit liability. We present a new tree-based algorithm that leverages mixed-integer optimization (MIO) techniques to generate interpretable and globally optimal tree-based clustering models. Our algorithm, Interpretable Clustering via Optimal Trees (ICOT), can incorporate various internal validation metrics, naturally determines the optimal number of clusters, and is able to account for mixed numeric and categorical data. It achieves comparable or superior performance on both synthetic and real world datasets when compared to k-means while offering significantly higher interpretability.

1 Introduction

In the era of Electronic Medical Records (EMR) and advanced health monitoring, the huge amount of data generated is too complex and voluminous to be analyzed by traditional methods [HDM15]. Unsupervised learning methods are able to transform this heterogeneous data into meaningful information for decision making [Elg+06]. However, if the outcome and the mechanism of clustering is not interpretable, the utility of the result is subject to doubt. Especially in a medical setting, where decision making can significantly impact individuals' disease trajectories, the characteristics of observations that are clustered in the same group need to be easily identifiable.

Considering the importance of cluster interpretability, there has been limited success in addressing the issue. The most popular method is the representation of a cluster of points by their centroid or by a set of distant points in the cluster which has been popular across various applications [Rad+04]. This works well when the clusters are compact or isotropic but fails when the clusters are elongated or non-isotropic. Another common approach is the visualization of clusters in a two-dimensional graph using principle component analysis (PCA) projections [Jol11; Rao64]. However, in reducing the dimensionality of the feature space, PCA obscures the relationship between the clusters and the original variables.

Tree-based methods such as CART [Bre+84] are a natural fit for problems that prioritize interpretability. A general approach involves running a traditional clustering method and assigning the resulting assignments as class labels. The data can then be fit using a classification tree and the decision paths leading to each cluster's leaves give insight into the differentiating features [HCE03]. While these trees give an explicit delineation of cluster attributes, the methods involve a two-step process of first building the clusters and subsequently identifying their differentiating features.

Motivated by the limitations of existing solutions to interpretable clustering, we propose a new tree-based machine learning algorithm, where interpretability is taken into consideration during cluster creation rather than considered as a later analysis step. Our method, called Interpretable Clustering via Optimal Classification Trees (ICOT), builds upon the algorithm of Optimal Classification Trees [BD17a] and extends it to the unsupervised setting. Our algorithm constructs a tree with a perspective of global optimality rather than taking a greedy approach. It is formulated as a mixed-integer optimization problem and can be solved using an iterative coordinate-descent approach that scales to larger problems.

We propose a non-heuristic unsupervised learning algorithm that solves the task at hand directly to optimality while providing the user with interpretable results based on the feature vectors. We use well-established validation criteria, such as the Silhouette Metric [Rou87] and Dunn Index [Dun74], as the algorithm's objective function taking into account both the inner-cluster density as well as the intra-cluster separation. Our technique renders the tuning of the tree's complexity redundant, making it easy to be used by medical researchers. The result is a tree that accounts for mixed numeric and categorical covariates and whose structure can be leveraged in a hierarchical way. We test the performance of our method compared to k-means in available datasets from the Fundamental Clustering Problems Suite (FCPS) and a real-world example from the Hubway system data repository. We demonstrate its superior performance in data with different levels of variance and compactness as well as its ability to provide us with interpretable clusters.

2 From Supervised to Unsupervised Learning

Interpretable Clustering via Optimal Trees (ICOT) builds trees through a modification of Optimal Classification Trees (OCT), a globally optimal tree-based algorithm [BD17b]. The resultant tree provides an explicit characterization of membership in a cluster, represented by a single leaf, through the path of feature splits. ICOT is formulated as a mixed-integer optimization problem (see Appendix), which creates a decision tree that minimizes a chosen loss function and assigns each observation to a leaf based on parallel feature splits. The algorithm is implemented using a coordinate-descent procedure which allows it to scale to much higher dimensions than directly solving the mixed-integer formulation, while still abiding by the same core principles.

ICOT initializes a greedy tree and then runs a local search procedure until the objective value, a cluster quality measure, converges. This process is repeated from many different starting greedy trees, generating many candidate clustering trees. The final tree is chosen as the one with the highest cluster quality score across all candidate trees. This single tree is returned as the output to the algorithm.

To form the initial greedy tree, we start with a single node and scan over potential splits on a randomly chosen feature. For each potential threshold for splitting observations into the lower and upper leaves, we compute the global score for the resultant assignment of the proposed split. After scanning through all thresholds, we choose the one that gives highest score and update the node to add the split if this score exceeds the global score of the current assignment. We perform the same search for each leaf that gets added to the tree, continuing until either the maximum tree depth is reached or no further improvement in our objective value is achieved through splitting a leaf.

Following the creation of the greedy tree, we begin the local search procedure. Nodes are visited in a randomly chosen order, and various modifications are considered. A "split" node (i.e. a node that is not a leaf) can be deleted, in which case it is replaced with either its lower or upper subtree, or a new split can be made at the node using a different feature and threshold. A leaf node can be further split and thus create two leaves. At each node, the algorithm finds the best possible change and then makes the proposed change only if it improves the objective from its current value. The algorithm terminates when the objective value converges.

2.1 Model Parameters

There are several user-defined inputs to the algorithm that give the user flexibility in their evaluation criterion and tree depth.

2.1.1 Cluster Quality Measure

The chosen loss function must consider the global assignment of observations to clusters, rather than each leaf in isolation as in OCT. This is due to the fact that the score of a clustering assignment depends on both the **compactness** of the observations within a single cluster, as well as its **separation** from observations in other clusters. Several internal validation metrics have been proposed to balance these two objectives [Liu+10]. The ICOT algorithm supports metrics that can be computed using the vector of observation cluster assignments and the pairwise Euclidean distance matrix for all observations. The distance matrix is independent of the cluster assignment and can thus be precomputed, lowering the computational intensity of the algorithm. Two common criteria, the Silhouette and Dunn Index scores, are outlined below.

Silhouette Metric The silhouette metric introduced by [Rou87] compares the distance from an observation to other observations in its cluster relative to the distance from the observation to other observations in the second closest cluster. The silhouette score for observation i is computed as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))} \tag{1}$$

where $a(i)$ is the average distance from observation i to the other points in its cluster, and $b(i)$ is the average distance from observation i to the points in the second closest cluster (i.e. $\min_k b(i,k)$ where $b(i,k)$ is the average distance of i to points in cluster k, minimized over all clusters k other than the cluster that point i is assigned to). This score ranges from -1 to 1, where a higher score is better. These individual scores can be averaged to reflect the quality of the global assignment.

Dunn Index The Dunn Index [Dun74] characterizes compactness as the maximum distance between observations in the same cluster, and separation as the minimum distance between two observations in different clusters. The metric is computed as the ratio of the minimum inter-cluster separation to the maximum intra-cluster distance. A high score is better, since it signifies that the distance between clusters is large relative to the distance between points within a cluster.

2.1.2 Tree Depth and Complexity

The natural balance between separation and compactness in the splitting process allows for a more exploratory approach to clustering. It eliminates the need for setting an explicit K parameter, which is typically required in both partitional and hierarchical clustering methods. The tree continues to split until further splits no longer improve the quality of the overall assignment, and so the final number of leaves represent the optimal number of clusters. This also eliminates the need for the complexity parameter, which is used in OCT to prevent oversplitting. The maximum depth can be used to impose an upper bound on a reasonable number of clusters if desired.

3 Results

3.1 Synthetic Datasets

We evaluated ICOT on the Fundamental Clustering Problems Suite datasets (FCPS) [Ult05], a standard set of synthetic datasets for unsupervised learning evaluation. Table 1 shows a comparison of ICOT, trained using the Dunn Index criterion, and k-means against the known labels. Both methods are able to recover the true cluster labels for three datasets.

The evaluation of cluster quality is highly dependent on the chosen metric. In two of the cases where ICOT does not capture the ground truth, namely the TwoDiamonds and Target datasets, the Dunn Index score is higher with ICOT's proposed clusters than on the ground truth. Additionally, it outperforms k-means on six of the datasets. This brings up the broader question of how to assess cluster quality; recovering known labels in synthetic data does not necessarily translate to meaningful cluster assignment. The ICOT validation criterion should be chosen in consideration of the desired cluster properties; the Dunn Index performs well at identifying clusters by geometric separation, while the Silhouette metric is often better at finding meaningful separation when accounting for the density of the data.

	Ground Truth	ICOT		K-Means	
Dataset	**Score**	**Score**	**Recovery?**	**Score**	**Recovery?**
Atom	.371	.137	No	.029	No
Hepta	1.076	.357	No	1.076	Yes
Lsun	.117	.117	Yes	.035	No
Target	.253	.362	No	.025	No
Tetra	.200	.200	Yes	.200	Yes
TwoDiamonds	.022	.044	No	.022	Yes
WingNut	.063	.063	Yes	.024	No

Table 1: Results of ICOT and k-means in comparison with ground truth on the FCPS datasets

Our method is unable to capture the ground truth when the underlying clusters are nonseparable with parallel splits (i.e. Atom, Hepta datasets). This is due to the fact that our method places hard constraints on an observation's cluster membership based on splits in feature values, whereas the other methods are less constrained. However, this trade-off allows for clear cluster definitions; thus we believe it is worthwhile within reason due to the necessity of interpretability and simple assignment rules in many settings.

3.2 A Real World Example

We provide an illustration of our method using data from the Offspring Cohort from the Framingham Heart Study, a large-scale longitudinal clinical study. The dataset comprises of 200 observations and 8 covariates (age, gender, presence of diabetes, levels of HDL, Systolic Blood Pressure, BMI and hematocrit). The ICOT algorithm creates 7 clusters corresponding to the leaves of the tree in Figure 1 and selects only four features to split on. The interpretable nature of ICOT allows us to understand the differentiating factors in these clusters. For example, we see a separation between younger women and men, and furthermore a different HDL threshold to distinguish among participants within each gender. Thus, we are able to clearly define the characteristics of each cluster.

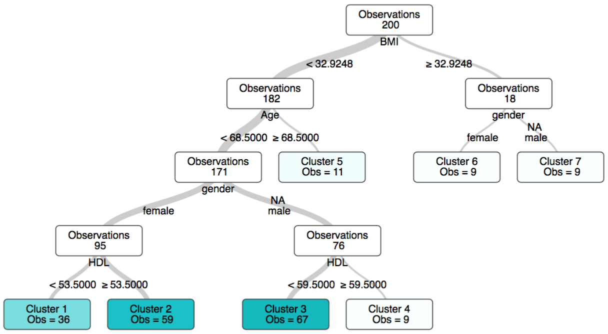

Figure 1: Visualization of the ICOT results for the Dunn index on Framingham Heart Study dataset.

4 Conclusions

We have introduced a new methodology of cluster creation that addresses the issue of cluster interpretability. Our method extends the framework of Optimal Classification Trees to an unsupervised learning setting, in which we build trees that provide explicit separations of the data on the original feature set. This makes it an ideal tool for exploratory data analysis since it reveals natural separations of the data with intuitive reasoning. We believe that our proposed clustering algorithm offers a promising alternative to existing methods, namely k-means and hierarchical clustering. Our early results suggest that we can recover clusters similar to k-means, but with the added advantages of interpretability and no prespecified cluster count. We hope to apply this method across various applications, particularly in healthcare, including grouping together similar patients, medical diagnoses, and others.

References

[Rao64] C. Radhakrishna Rao. "The Use and Interpretation of Principal Component Analysis in Applied Research". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26.4 (1964), pp. 329–358. ISSN: 0581572X. URL: http://www.jstor.org/stable/25049339.

[Dun74] J C Dunn. "Well-Separated Clusters and Optimal Fuzzy Partitions". In: *Journal of Cybernetics* 4.1 (1974), pp. 95–104. DOI: 10.1080/01969727408546059org/10.1080/01969727408546059. URL: http://www.tandfonline.com/action/journalInformation?journalCode=ucbs19.

[Bre+84] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.

[Rou87] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. URL: https://ac.els-cdn.com/0377042787901257/1-s2.0-0377042787901257-main.pdf?%7B%5C_%7Dtid=0425bd02-6f15-4f02-9634-aabba2d8c046%7B%5C&%7Dacdnat=1532793956%7B%5C_%7D.

[HCE03] T P Hancock, D H Coomans, and Y L Everingham. "Supervised Hierarchical Clustering Using CART". In: *Proceedings of MODSIM 2003 International Congress on Modelling and Simulation*. Townsville, QLD, Australia, 2003, pp. 1880–1885. ISBN: 978-1-74052-098-0. URL: https://www.mssanz.org.au/MODSIM03/Volume%7B%5C_%7D04/C07/06%7B%5C_%7DHancock.pdf.

[Rad+04] Dragomir R. Radev et al. "Centroid-based summarization of multiple documents". In: *Information Processing Management* 40.6 (2004), pp. 919–938. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2003.10.006. URL: http://www.sciencedirect.com/science/article/pii/S0306457303000955.

[Ult05] A Ultsch. *Fundamental clustering problems suite (fcps)*. Tech. rep. University of Marburg, 2005. URL: https://github.com/Mthrun/FCPS/.

[Elg+06] Haytham Elghazel et al. "A New Clustering Approach for Symbolic Data and Its Validation: Application to the Healthcare Data". In: *Foundations of Intelligent Systems*. Ed. by Floriana Esposito et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 473–482. ISBN: 978-3-540-45766-4.

[Liu+10] Yanchi Liu et al. "Understanding of internal clustering validation measures". In: *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE. 2010, pp. 911–916.

[Jol11] Ian Jolliffe. "Principal component analysis". In: *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[HDM15] Ramzi A. Haraty, Mohamad Dimishkieh, and Mehedi Masud. "An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data". In: *International Journal of Distributed Sensor Networks* 11.6 (2015), p. 615740. DOI: 10.1155/2015/615740. eprint: https://doi.org/10.1155/2015/615740. URL: https://doi.org/10.1155/2015/615740.

[BD17a] D. Bertsimas and J. Dunn. "Optimal Trees". In: *Machine Learning* 106.7 (2017), pp. 1039–1082.

[BD17b] Dimitris Bertsimas and Jack Dunn. "Optimal classification trees". In: *Machine Learning* (2017), pp. 1–44.

5 Appendix

5.1 Mixed Integer Formulation

The OCT algorithm formulates tree construction as a MIO which us to define a single problem, as opposed to the traditional recursive, top-down methods that must consider each of the tree decisions in isolation. It allows us to consider the full impact of the decisions being made at the top of the tree, rather than simply making a series of locally optimal decisions, avoiding the need for pruning and impurity measures.

The Optimal Classification Trees (OCT) framework can be modified to address an unsupervised learning setting. We present changes in the original MIO formulation of OCT to be able to partition the data space into distinct clusters following the same structure and notation as in [BD17b]. There are two primary modifications in our model:

1. The objective function is comprised solely by the chosen cluster quality criterion, such as the silhouette score, and does not include any penalty for the tree complexity.
2. Each leaf of the tree is equivalent to a cluster. Observations in different leaves are not allowed to belong to the same cluster.

Given a tree object, we will index its nodes by $t = 1, \dots, T$. We use the notation $p(t)$ to refer to the parent node of node t, and $A(t)$ to denote the set of ancestors of node t. We also define $A_L(t)$ as the set of ancestors of t whose left branch has been followed on the path from the root node to t, and similarly $A_R(t)$ is the set of the right-branch ancestors.

The nodes in the tree are divided into two sets:

- Branch nodes: Nodes $t \in \mathcal{T_B}$ apply a split of the form $a^x < b$. All the points that satisfy the split follow the left branch in the tree and those that do not follow the right branch.
- Leaf nodes: Nodes $t \in \mathcal{T_L}$ formulate a cluster for all the points that fall into the leaf node.

As in the OCT formulation, we define the split applied at node $t \in \mathcal{T_B}$ with variables $\mathbf{a}_t \in \mathbb{R}^p$ and $b_t \in \mathbb{R}$. The vector $\mathbf{a}_t$ indicates which variable is chosen for the split, meaning that $a_{jt} = 1$ for the variable j used at node t. b_t gives the threshold for the split, which is between $[0, 1]$ after normalization of the feature vector. Together, these form the constraint $\mathbf{a}_t^T x < b_t$. The indicator variables d_t are set to 1 for branch nodes and 0 for leaf nodes. Using the above variables, we introduce the following constraints that allows us to model the tree structure (for a detailed analysis of the constraints, see [BD17b]):

$$\sum_{j=1}^{p} a_{jt} = d_t, \ \forall t \in \mathcal{T_B}, \tag{2}$$

$$0 \le b_t \le d_t, \ \forall t \in \mathcal{T_B}, \tag{3}$$

$$a_{jt} \in \{0,1\}, \ j = 1, \dots, p, \quad \forall t \in \mathcal{T_B}, \tag{4}$$

$$d_t \le d_{p(t)}, \ \forall t \in \mathcal{T_B} \backslash \{1\}, \tag{5}$$

Next we present the corresponding constraints that track the allocation of points to leaves. For this purpose, we introduce the indicator variables $z_{it} = 1\{x_i$ is in node $t\}$ and $l_t = 1\{$leaf t contains any points$\}$. We let N_{min} be a constant that defines the minimum number of observations required in each leaf. We apply the following constraints as in OCT:

$$\sum_{t \in \mathcal{T_L}} z_{it} = 1, \ i = 1, \dots, n, \tag{6}$$

$$z_{it} \le l_t, \ \forall t \in \mathcal{T_L}, \tag{7}$$

$$\sum_{i=1}^{n} z_{it} \ge N_{min} l_t, \ \forall t \in \mathcal{T_L} \tag{8}$$

Next we present the set of constraints that enforce the splits that are required by the structure of the tree when assigning points to leaves. We want to enforce a strict inequality for points going to the lower leaf. To accomplish this, we define the vector $\epsilon \in \mathbb{R}^p$ as the smallest separation between two observations in each dimension p, and ϵ_{max} as the maximum over this vector. The split can then be enforced using the following constraints:

$$\begin{align} a^{x}_{mi} \geq b_t - (1 - z_{it}),\ i = 1, \dots, n, \quad \forall t \in \mathcal{T_B}, \quad \forall m \in A_R(t) \tag{9} \\ a^{(}_{m} x_i + \epsilon) \leq b_t + (1 + \epsilon_{max})(1 - z_{it}),\ i = 1, \dots, n, \quad \forall t \in \mathcal{T_B}, \quad \forall m \in A_L(t) \tag{10} \\ \tag{11} \end{align}$$

The objective of the new formulation is to maximize the silhouette score S of the overall partition. The silhouette score quantifies the difference in separation between a point and points in its cluster, vs. the separation between that point and points in the second closest cluster.

Let d_{ij} be the distance (i.e. Euclidean) of observation i from observation j. We define K_t to be number of points assigned assigned to cluster t.

$$K_t = \sum_{i=1}^{n} z_{it} \forall t \in \mathcal{T_L} \tag{12}$$

We define $c_i t$ to be the average distance of observation i from cluster t:

$$c_{it} = \frac{1}{K_t} \sum_{j=1}^{n} d_{ij} z_{jt},\ i = 1, \dots, n,\ \forall t \in \mathcal{T_L}. \tag{13}$$

We define r_i to be the average distance of observation i from all the points assigned in the same cluster:

$$r_i = \sum_{\forall t \in \mathcal{T_L}} c_{it} z_{it},\ i = 1, \dots, n. \tag{14}$$

We then let w_{it} denote the minimum average distance of observation i from all the points in cluster t where i does not belong to cluster t.

$$q_i \leq c_{it}(1 - z_{it}) + M z_{it},\ i = 1, \dots, n,\ \forall t \in \mathcal{T_L}. \tag{15}$$

Finally, to define the silhouette score of observation i, we will need the maximum value between r_i and q_i which normalizes the metric.

$$m_i \geq r_i,\ i = 1, \dots, n. \tag{16}$$

$$m_i \geq q_i,\ i = 1, \dots, n. \tag{17}$$

The silhouette score for each observation is computed as $s(i)$ and the overall Silhouette score for the clustering assignment is then the average overall all the silhouette scores from the training population:

$$s_i = \frac{q_i - r_i}{m_i},\ i = 1, \dots, n. \tag{18}$$

$$S = \frac{1}{n} \sum_{i=1}^{n} s_i. \tag{19}$$