Capturing the Essence: Partial Explanations for Interpretable Clustering

Kevin Quinn[1] (✉), Evimaria Terzi[1], and Heikki Mannila[2]

[1] Boston University quinnk@bu.edu
[2] Aalto University

Abstract. Recent work on interpretable clustering focuses on forming a partition of the input data such that each cluster is associated with a logical, conjunctive rule description, and where every point must belong to one cluster. We build upon previous work by lifting the requirement that our cluster descriptions should characterize the entire dataset. Rather, we show that by ignoring outliers and boundary points lying between multiple clusters, we produce shorter rule descriptions with improved cluster quality. To that end, we define the Partial Interpretable Clustering problem, taking a dataset as input and producing a rule-based clustering for a subset of data points as output. We design a framework which builds concise, conjunctive rules with decision trees, collecting them to form more generally structured decision sets. For any given decision set, we then introduce efficient algorithms which leverage techniques from submodular optimization and outlier detection in order to identify the subset of the data points we choose to describe. Our experiments with real datasets demonstrate the efficacy of our methods in creating cohesive clusters – in terms of distance or cost – which are amenable to short and precise descriptions – in terms of conjunctive rule length.

Keywords: Interpretable Clustering · Explainable AI · Decision trees · Decision sets · Climate Data

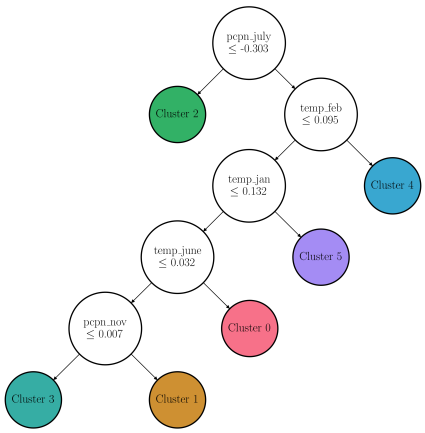
1 Introduction

Interpretable machine learning models use simple, logical rules as the core of their decision-making process [25,26]. When rules are concise and easy to understand, they allow for both strong, critical data analysis and deeper insight into a model's ability to make accurate and sensible predictions. Unsupervised clustering problems are one domain in which interpretability has recently taken a strong hold [15]; it has been shown [9,11,23] that binary decision trees can be used to recursively split data points based on simple inequalities for individual features, while still maintaining competitive cost performance.

Our work departs from previous approaches to interpretable clustering [1,23] by adopting the perspective that an interpretable clustering model need not *cover* or explain every point in its input dataset. Furthermore, we produce clusterings for which clusters may *overlap* or share common data points. With such

relaxations, we open the door for cluster descriptions with considerably different logical structure, producing rules which are shorter and easier to parse, and often improve the clustering performance on the subsets of points that remain covered.

Specifically, we design clustering models in the form of *decision sets*. Whereas a tree makes recursive decisions which are forced to account for the entire dataset all at once, a decision set makes local decisions as an unstructured list of rules in the form: *If x then cluster y.* User studies have previously shown that decision sets are more immediately understood by readers [18], however we also notice an apparent advantage in the *rule length* i.e., the maximum number of logical conditions used to distinguish any cluster. As an extreme example, consider a $k \times k$ grid in two dimensions, where each grid cell contains a distinct cluster. Any decision tree must have a depth or rule length of at least $\Omega(\log_2(k))$ to distinguish each of the k^2 clusters. Whereas a decision set may have a constant rule length of four: for each cluster, an inequality for each of its four side boundaries. Similarly, consider a dataset with k clusters for which points in cluster i contain ones in feature i and zeros everywhere else. In this case, any decision tree must have depth $k-1$, since we are constrained to sequentially distinguishing one cluster from the rest. A decision set on the other hand may simply use one logical condition for each cluster: *if $x_i > 0$ then cluster i.*



(a) `IMM` Decision Tree

If (pcpn_apr ≤ -0.208) & (pcpn_july > -0.246)
Then cluster 0

If (temp_jan ≤ 0.04) & (pcpn_mar ≤ -0.186)
Then cluster 1

If (pcpn_july ≤ -0.275) & (temp_may > 0.023)
Then cluster 2

If (temp_july ≤ 0.01) & (temp_jan ≤ 0.134)
Then cluster 3

If (temp_oct > 0.047) & (temp_feb > 0.064)
Then cluster 4

If (temp_apr ≤ -0.011) & (temp_nov > 0.034)
Then cluster 5

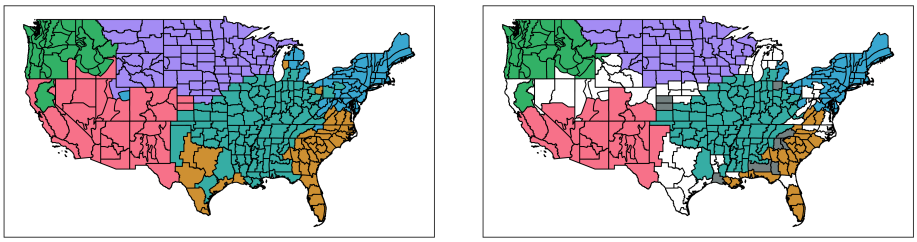
(b) `Forest-Depth-2` Decision Set

Fig. 1: **Decision Rules** separating 6 clusters for the *Climate* dataset. We compare the readability of a decision tree clustering (left), as compared to a decision set (right) which satisfies 80% of data points.

In this paper, we introduce algorithms for both building large sets of concise rules and selecting size-k subsets of them to form a *decision set* clustering. With a random forest approach, we train a collection of trees in which only a

single chosen cluster is distinguished from the rest, allowing for brief description. This is complemented by a submodular pruning objective and an efficient greedy heuristic designed to maximize coverage while minimizing overlap. With experimental evidence, we show that our algorithms exhibit efficient clustering performance, measured with a novel definition of clustering *distortion* on the subset of covered points.

To showcase our method's interpretability and analytical strengths, we display qualitative evidence with a novel analysis of a *Climate* change dataset, containing information on percent change in temperature and precipitation levels for various locations in the continental US [24,28] (see section 5.2 for a complete description). Figure 1 compares a decision tree model to that of a decision set. Whereas the decision tree requires a depth or rule length of five logical conditions to describe the two clusters at its furthest leaf nodes, the decision set is built from rules with a consistent length of two. Moreover, Figure 2 shows geographical evidence that the resulting clustering is comparable. Although some locations are overlapped or left uncovered, we notice that these locations are often found near the boundaries between clusters. In fact, our experiments highlight that our method is effective in removing boundary points for which cluster membership is indistinct, allowing for models with improved clustering performance on the remaining set of points.[3]



(a) `IMM` Decision Tree (b) `Forest-Depth-2` Decision Set

Fig. 2: **Clustered Maps** for the *Climate* dataset. We compare a complete partition clustering with a decision tree (left) to a partial decision set clustering satisfying 80% of data points (right). Samples which are uncovered are left uncolored and those with overlapping membership are colored in grey.

2 Related Work

With a rise in popularity of interpretable machine learning [25,26], more attention is being given to its adoption into the toolbox of clustering algorithms.

[3] All code and data for reproducing our experiments is anonymously available here.

Beginning with work on the novel iterative mistake minimization (`IMM`) algorithm [23], many approaches have proposed new and interesting variations on decision tree clusterings [9,10]. We provide a complete description of `IMM` within our experimental section and direct reader attention to a more complete survey [15] for its surrounding variations.

Our most direct relation from this line of work comes from Bandyapadhyay et al. [1], who study decision tree explanations in the context of outlier removal. Inspired by outlier removal for problems such as `KMeans` [5,21], this study seeks to determine an minimal number of points, s, to be removed in order to exactly replicate a given reference clustering (obtained with `KMeans` or similar) in the form of a decision tree. To do so they design an efficient approximation algorithm which removes at most $s(k-1)$ misclassified points during creation of their tree. Since s is an unknown constant, however, this offers no practical control over the number of points actually removed. Their work also analyzes an algorithm with more controlled removal, but for the purpose of complexity results rather than as an efficient algorithm to be used in practice.

Other related perspectives have emphasized the need for better explainability within decision trees [16,17], introducing penalties for both rule length (tree depth) as well as the total number of nodes within a tree. While our study is similarly concerned with both outlier removal and improved explainability, we build upon the aforementioned work with the use of unstructured decision sets, freeing us from the constraint of needing a rule length of at least $\log_2 k$ to sequentially separate k clusters. Furthermore our work highlights the relationship between objectives of coverage, cost, and rule length; we observe tradeoffs between large coverage with small rule length and clustering cost performance. We argue that as a whole, explainable clustering with unstructured sets of rules is an approach which remains under-studied. Recent work has focused on creating rectangular or polyhedral rule descriptions [6,19,30], others have studied association rules in the context of graph clustering [27], and there is also a longer history of unstructured fuzzy rule descriptions in which points are given degrees of associated cluster membership [12,22]. To our knowledge, however, none have done so while simultaneously studying a notion of partial clustering or coverage.

We also note that similar rule selection frameworks have recently been considered for explainable clustering problems [4,13], in which descriptions are chosen from a larger set of rules in order to create an explanatory model for a reference clustering. As in our own work, objectives in these settings are designed to capture the tension found in maintaining accurate cluster coverage while minimizing overlap. However, they define and solve their problems using complex mixed-integer or conditional programs. Furthermore, their rule-generation processes are left underspecified [4], or are only applicable for categorical data types [13]. We argue that our framework introduces both improved methods for rule generation, as well as a simple and considerably more efficient (polynomial time) algorithm for rule selection.

Our study has also taken significant inspiration from work outside the immediate realm of interpretable clustering. Specifically, we build directly upon

work which uses decision sets in the context supervised learning problems with categorical data [18,31]. For these settings, authors have designed useful submodular objectives which encourage large-coverage selections of accurate and interpretable rules from within a larger decision set collection. Our work extends theirs by introducing a novel submodular objective which is efficiently solved with the recently-proposed distorted greedy algorithm [14]. Moreover, by considering the unsupervised setting, we design a rule-generation process for real-valued data, and also introduce the notion of partial clustering cost and clustering distortion.

3 Notation

3.1 Clustering

Let $X = \{x_1, x_2, ..., x_n\}$ be a dataset with n entries $x_i \in \mathbb{R}^d$. For a single point $x \in X$, let x_i be its value for the ith feature. We say that a *clustering* $\mathcal{C}$ is a size k set of data subsets or clusters $C_i \subseteq X$. Importantly, we depart from standard definitions by allowing clusters to share common data points, as well as by allowing for points which do not belong to any cluster. For any clustering let $X_{\mathcal{C}} = \bigcup_{C_i \in \mathcal{C}} C_i$ be its set of *covered* points, which belong to at least one cluster. For any subset of points $Y \subseteq X$ let its partial clustering $\mathcal{C}(Y)$ be the condensed set of clusters $C_i \cap Y$ for all $C_i \in \mathcal{C}$. We say the size of a partial clustering $|\mathcal{C}(Y)|$ is its number of non-empty cluster sets. For a single point, x, if $|\mathcal{C}(x)| > 1$ then it is said to *overlapping*, since it must belong to multiple clusters.

For any cluster C_i, its corresponding center, u_i, is defined as mean of its data points. We use $\mathcal{U}$ when referring to the complete set of centers for each C_i. To evaluate clusterings which may have overlapping cluster membership or uncovered points we consider the following normalized variation of the classic sum of squared errors cost.

Definition 1 (Clustering Cost). *For a clustering $\mathcal{C}$,*

$$\text{cost}(\mathcal{C}) := \frac{1}{X_{\mathcal{C}}} \sum_{i=1}^{k} \sum_{x \in C_i} \frac{\|x - u_i\|_2^2}{|\mathcal{C}(x)|}. \tag{1}$$

Finally, throughout our study we will focus on making comparisons between different clusterings. Since clusterings may cover different sets of data points, however, even a normalized cost comparison will fail to capture a meaningful relationship. For example, one clustering may see a smaller cost simply because it ignores a distant data point which another does not, rather than because it actually improves upon the underlying model. Therefore, we compute a cost ratio between clusterings for a given set of commonly covered points.

Definition 2 (Distortion). *For clusterings $\mathcal{C}$, $\mathcal{C}'$ and a subset of data Y such that $Y \subseteq X_{\mathcal{C}}$ and $Y \subseteq X_{\mathcal{C}'}$,*

$$\text{distortion}(\mathcal{C}, \mathcal{C}', Y) := \frac{\text{cost}(\mathcal{C}(Y))}{\text{cost}(\mathcal{C}'(Y))}. \tag{2}$$

For a set of points Y which isn't fully contained by either clustering, we say that the distortion is undefined or infinite.

3.2 Predicates and Rules:

A predicate $r : \mathbb{R}^d \to \{\text{True}, \text{False}\}$ is a boolean function which returns the evaluation of a point x on a given logical condition. For our purposes, we restrict our attention to simple axis aligned conditions of the form $x_i \leq \tau$. A rule R is a boolean function whose value for point x is defined by a logical composition of the values $r_1(x), \ldots, r_w(x)$ for x of some predicates $r_1, \ldots, r_w$. We focus our attention on rules which are conjunctions of predicates, and therefore use only logical $\wedge$ (AND) operators. We use $|R|$ to denote the length of the rule, or the number of predicates r conjoined within its formula. We also define X_R to be the subset containing all points $x \in X$ for which $R(x)$ is True. Likewise a rule set $\mathcal{S}$ is taken to be a collection of rules R, and we use $X_{\mathcal{S}} := \bigcup_{R \in \mathcal{S}} X_R$ to denote the set of all points from X covered by at least one rule in $\mathcal{S}$.

3.3 Decision Sets

Let a decision set $\mathcal{D}$ be a pair $(\mathcal{S}, \psi)$, for which a rule assignment function $\psi : \mathcal{S} \to \mathbb{N}$ maps each rule $R \in \mathcal{S}$ to a single cluster label. Notice that any decision set induces a very natural clustering upon a given dataset X. Specifically, let cluster $C_i = \bigcup_{R \in \mathcal{S}, \psi(R)=i} X_R$ be the set of all points covered by a rule with label i. We refer to an induced decision set clustering as $\mathcal{C}_{\mathcal{D}}$. Finally, we note that a decision tree $\mathcal{T}$ is a specialized form of a decision set for which rules $\mathcal{S}$ are recursively structured in the form of a binary tree. Each rule $R \in \mathcal{S}$ may therefore be thought of as a conjunction of conditions found on a path from the root to a given leaf.

4 Partial Interpretable Clustering

We now introduce a partial clustering problem, which is studied throughout the rest of our paper. Tied to the idea that an interpretable clustering need not describe every instance, we search for low-cost decision set clusterings that satisfy given coverage constraints. In doing so we adopt an approach which builds upon a given reference clustering $\mathcal{C}_{\text{ref}}$, found with `KMeans`. Specifically, we focus our attention upon a relative comparison which considers the effect of using a decision set clustering to model or characterize a set of points shared with the reference.

Problem 1. Partial Interpretable Clustering (PIC)
Given a dataset X with size n and a reference clustering $\mathcal{C}_{\text{ref}}$, find a decision set $\mathcal{D}$ and subset of data $Y \subseteq X$ with size $|Y| \geq \tau n$ for $\tau \in [0, 1]$ such that the induced clustering $\mathcal{C}_{\mathcal{D}}$ upon Y satisfies minimal distortion,

$$\underset{\mathcal{D},\, Y \; s.t. \; |Y| \geq \tau n}{\arg\min} \quad \text{distortion}(\mathcal{C}_{\mathcal{D}}, \mathcal{C}_{\text{ref}}, Y). \tag{3}$$

The remainder of this section describes the heuristic strategies we consider for solving the PIC problem. Our strategies are designed assuming access to methods for fitting an appropriate decision set $\mathcal{D}$ to a reference clustering, and we describe our methods for doing so in the next section. As a starting point, we note that the algorithm of Bandyapadhyay et al. [1] (described in section 5.1) produces a solution with ratio exactly 1, *if* it happens to satisfy the coverage requirement. However, since the algorithm offers no control over how many points are removed, we do not consider it to be a robust solution.

4.1 Removing Outliers

A naive strategy may start by removing a subset $Z \subseteq X$ of outliers from $\mathcal{C}_{\text{ref}}$, and then retrain a decision set on the partial clustering $\mathcal{C}_{\text{ref}}(Y)$, where $Y = X \setminus Z$. While it is not immediately clear that doing so would improve the cost ratio, it does leave the door open for fine-grained control over the number of points removed. Throughout our work, we consider outlier points to be those that lie on the boundaries between clusters and are therfore difficult to describe with axis-aligned rules. More specifically, for a point x and a set of centers $\mathcal{U}$, let u^* be the center smallest distance to x and u' be the center with the next smallest distance. The distance ratio is then defined as follows.

Definition 3 (Distance Ratio). *For a given point $x \in X$ and set of representative centers $\mathcal{U}$,*

$$\text{distance-ratio}(x, \mathcal{U}) := \frac{\|x - u'\|_2}{\|x - u^*\|_2}. \tag{4}$$

Intuitively if a data point lies between two clusters, or is otherwise sufficiently far from every cluster, the ratio will be close 1. Therefore, when using an outlier removal strategy, we rank points by their distance ratio to centers from $\mathcal{C}_{\text{ref}}$, and remove a $1 - \tau$ fraction of the smallest.

4.2 Selecting Rules

Alternatively, we consider a method in which k rules are selected from a trained decision set $\mathcal{D} = (\mathcal{L}, \psi)$, where a set of m rules $\mathcal{L}$ satisfies $m \geq k$. To do so, we focus our attention upon rules which best describe the reference clustering, and which must therefore also produce smaller cost ratios. In particular, we state an objective which accounts for both accurate coverage of the reference clusters, as well as minimal overlap among the set of selected rules. First, we compute the number of cluster points which have been covered by a rule with the same label by the function

$$g(\mathcal{S}) = \sum_{i=1}^{k} \left| \bigcup_{R \in \mathcal{S}, \psi(R) = i} X_R \cap C_i \right|. \tag{5}$$

Next, since we would also like to ensure that the rules selected to represent a cluster have minimal overlap with data points from other clusters, we pair the previous objective with an overlap penalty and state the following claim:

$$c(\mathcal{S}) = \sum_{i=1}^{k} \sum_{R \in S, \psi(R)=i} |X_R \cap (X \setminus C_i)| \tag{6}$$

Claim. The function $g(\mathcal{S})$ is a monotone, submodular function and $c(\mathcal{S})$ is a monotone, modular function.

To see why, notice that $g(\mathcal{S})$ is simply a sum of k coverage functions – which are not unlike standard set-cover objectives – and that $c(\mathcal{S})$ is a sum in which rules R are always penalized independently (there is no interaction between rules in the sum). It follows that a combined objective,

$$f(\mathcal{S}) := g(\mathcal{S}) - \lambda c(\mathcal{S}), \tag{7}$$

is a submodular-modular function, for which maximization would mean both large accurate coverage and small overlap. Such a problem: $\max_{\mathcal{S} \subseteq \mathcal{L}, |\mathcal{S}|=k} f(\mathcal{S})$ is, however, NP-hard. We therefore rely upon results from Harshaw et al. [14], who show that a $(1 - \frac{1}{e})$ approximation is achievable. Specifically, their distorted greedy algorithm iteratively adds new rules R to a developing set $\mathcal{S}$ by greedily taking those which maximize a weighted sum of over items $g(\mathcal{S} \cup R) - g(\mathcal{S})$ and $\lambda c(R)$. In early iterations the algorithm favors rules with large coverage, but gives more consideration to overlap as the size of the solution set increases. Since each round computes the marginal gain of all potential rules and we repeat the process k times, the runtime may be expressed as $\mathcal{O}(kmn)$.

Also included within Equation (7) is a tuning parameter λ which allows for a controlled trade-off between accurate coverage and overlap, with $\lambda = 0$ admitting solutions that may cover arbitrarily large portions of the input space. To ensure that the coverage requirement τ is satisfied for PIC, we therefore search over a given range Λ of λ values, allowing for large overlap by using smaller λ values if necessary. The price of doing so, however, is increased cost for the clustering induced by the selected set of rules $\mathcal{S}$ and their assignment ψ. Whenever we find multiple solutions which satisfy the coverage requirements, we therefore favor those that produce clusterings with smaller cost.

5 Experimental Results

5.1 Algorithms

In our experiments we compare against the following baseline algorithms for producing clusterings and decision trees. For all decision trees considered, data is recursively split in the form of feature threshold pairs (j, θ) at every non-leaf node. First, `KMeans` is used to produce reference clusterings $\mathcal{C}_{\text{ref}}$, and we use a standard implementation with a `KMeans` ++ initialization method. Next, `IMM` or iterative mistake minimization [23] is an interpretable method which builds a decision tree utilizing centers u from $\mathcal{C}_{\text{ref}}$. Starting at the root, both data points and centers are recursively split until the tree has exactly k leaves (each

containing a single cluster center). Every leaf (rule) is assigned the label of the cluster whose representative satisfies its conditions. Finally, we also consider the explainable clustering algorithm designed by `Bandyapadhyay et al.` [1]. In their algorithm a decision tree brings outlier removal into its training process by removing points which become separated from the majority of their cluster after any new split. Splitting conditions are therefore greedily chosen to minimize the number of removals.

These are then compared to the the strategies described in section 4. First, we use `IMM-outliers` as an algorithm which removes outlier points using distances to centers in $\mathcal{C}_{\text{ref}}$ and trains an `IMM` tree upon the remaining set of data.

Our main strategy, however, is the interpretable `Forest` clustering algorithm. For it we start by creating a random forest in which canonical decision trees are each trained to distinguish a single, uniform random cluster with label i in a one-versus-all fashion. In other words, each tree solves a binary classification problem where training points are given label 1 if they belong to cluster i and label 0 otherwise. Note that we use reference cluster labels to do so and, therefore, $\mathcal{C}_{\text{ref}}$ must be a standard, full partition clustering. Leaf node rules R are collected to form a general purpose decision set by taking those with a predicted label of 1, and assigning them to the cluster being distinguished, so that $\psi(R) = i$. Each tree is trained upon a random 75% portion of the original data set, and is constrained to use at most a depth (rule-length) uniformly chosen from $[1, h]$. We describe variations of the algorithm with maximum allowed depth h as `Forest-Depth-h`. All experiments are performed using 1000 trained trees, and the entire process is repeated and averaged over 1000 random trials.

Each time a forest decision set has been formed, we select a size k subset using the distorted greedy procedure outlined in section 4.2. To choose values of λ we find that an inexpensive grid search over 50 values in the range of $\Lambda = [0, 5]$ is sufficient to explore the space of solutions for our experiments.

5.2 Datasets

To test our algorithms, we make use of the following datasets. Unless otherwise specified, we take the number of clusters to be the number of unique classification labels associated with the dataset. Our first three real valued datasets are preprocessed using standard scaling, and the remaining image datasets are normalized to the range of $[0, 1]$.

The ***Climate*** ($n = 344$, $d = 24$, $k = 6$) dataset is collected and maintained by the National Oceanic and Atmospheric Administration [24,28]. It is comprised of 344 climate division locations, which partition the continental US into small zones of similar climate. For each of 12 months in a year, measurements for climate divisions are given by both a percent change in temperature (`temp`) and percent change in precipitation (`pcpn`) levels (giving 24 total features). Percent change is computed as the relative difference between average values for temperature and precipitation during period of 2013-2023, and averages from a historical period of 1900-2000. The number of clusters is chosen using an elbow heuristic on an analysis of `KMeans` cost. The ***Anuran*** ($n = 7195$, $d = 22$,

$k = 10$) dataset contains samples of anuran frog call recordings, measured over a frequency spectrum with 22 variables, and which are classified into ten different sub-species [7]. The ***Cover*** ($n = 145253$, $d = 10$, $k = 7$) dataset characterizes forest cover types for samples taken from wilderness areas in Colorado [2]. All categorical features were removed and the number of samples is a random 25% of the original dataset. Finally, the ***Digits*** ($n = 1797$, $d = 64$, $k = 10$)[20], ***Mnist*** ($n = 17500$, $d = 784$, $k = 10$) [8], and ***Fashion*** ($n = 17500$, $d = 784$, $k = 10$) [29] datasets are used as standard datasets for evaluation. For *Mnist* and *Fashion*, the number of samples is a random 25% of the original data.

5.3 Distortion with Increasing Coverage

Our central experiment studies the dynamic relationship between the minimum coverage requirement, τ, and clustering distortion for each of our algorithms. Figure 3 showcases the result of recomputing and measuring our algorithms as coverage requirements are incremented from 50-100%. In each plot, we display the distortion of the `IMM-outliers` and `Forest` algorithms relative to a single, common `KMeans` reference clustering. At each coverage step, the `IMM-outliers` distortion is computed upon its subset of remaining non-outlier points, and for all `Forest` methods we use the set $X_{\mathcal{S}}$ of points covered by rules selected with the greedy procedure. For each plot we also include an `IMM` baseline distortion, which is computed relative to the common reference clustering and is also taken over the same subset of data points used by the algorithm it compares against. Finally we note that three variations of `Forest` were included in the experiment, which sequentially allow for their trees to take maximum depths of 2,3, and 4.

We begin with the observation that the `IMM-outliers` algorithm does poorly to improve upon its `IMM` baseline. While we notice improvements within regions of low coverage requirement for the *Climate*, *Anuran*, and *Digits* datasets, the pattern is inconsistent and quickly diminished as τ increases. We therefore conclude that PIC is a problem which requires more careful algorithm design. For the displayed `Forest` algorithms, we observe a strong pattern in which distortion is often close to the reference value of 1 until around 90% coverage, where it typically sees drastic increase. We take this as evidence that the algorithms are efficient in covering easily explained parts of the data, but for large coverage requirements, are forced to select rules with a high degree of overlap. Note that complete coverage is not a guarantee and the displayed lines are cut off once the selection process cannot find anything larger. This is dependent upon the generated set of rules, and for a dataset such as *Mnist*, our rules appear to be discerning enough to only produce low coverage solutions. Among all `Forest` experiments, we also observe that the distortion performance is significantly improved by increasing rule length. The most dramatic difference is seen for the covertype dataset, which performs very poorly with an initial depth of 2. We argue that efficient performance is not always a given with small-length rules, especially for clusterings which are dependent upon the full feature space.

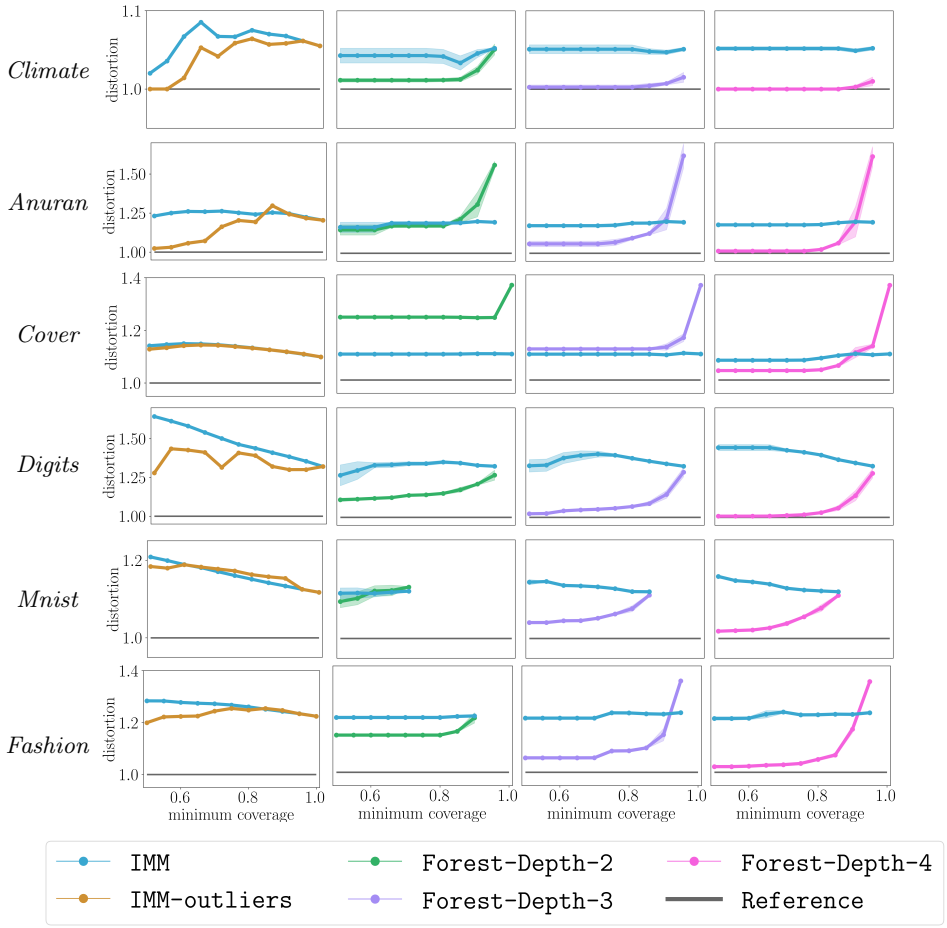


Fig. 3: **Distortion with increasing coverage.** Columns display distortion (y-axis) as a function of minimum required coverage, τ (x-axis) for each of the decision set algorithms respectively. Distortion is computed relative to a `KMeans` reference and is compared to an `IMM` baseline in each plot. Rows display results from each experimental dataset.

5.4 Interpretability Evaluation

Motivated by the goal of producing highly interpretable rules, we also show evidence that our `Forest` algorithms do well to outperform baseline decision tree algorithms in terms of rule length, while still maintaing small overlap and reasonably large coverage. Taking inspiration from [17,18], we evaluate our algorithms based upon the following interpretability measurements, and report comparisons to our baselines in Table 1. We observe that even the most complex `Forest-Depth-4` often gives significant improvement to both maximum and weighted rule length. We also notice that average overlap tends to be small, with a large majority of data points belonging to a single cluster in all methods. We note that for the algorithm of `Bandyapadhyay et al.`, without control over coverage we find scenarios such as the *Digits* or *Fashion* datasets for which its default cover is significantly lower than that of `Forest`, even though our previous experiment shows `Forest` maintains low distortion.

Definition 4 (Maximum and Weighted-Average Rule Length). *For a dataset X and ruleset $\mathcal{S}$, we measure the maximum length of all its rules as* $\text{max-length}(\mathcal{S}) := \max_{R\in\mathcal{S}} |R|$. *With a coverage sum,* $q(\mathcal{S}) := \sum_{R\in\mathcal{S}} |X_R|$, *we also say that*

$$\text{weighted-length}(X, \mathcal{S}) := \frac{1}{q(\mathcal{S})} \sum_{R\in\mathcal{S}} |X_R||R|. \tag{8}$$

Definition 5 (Coverage and Overlap). *For a given dataset X with size n and clustering $\mathcal{C}$, let the fraction of points belonging to at least one cluster be* $\text{coverage}(\mathcal{C}) := \frac{|X_\mathcal{C}|}{n}$. *Likewise, let the average cluster membership be evaluated as*

$$\text{overlap}(\mathcal{C}) := \frac{1}{|X_\mathcal{C}|} \sum_{x\in X_\mathcal{C}} |\mathcal{C}(x)|. \tag{9}$$

5.5 Characterizing Uncovered and Overlapped Points

Finally, we give evidence supporting the conclusion that data points which are left uncovered, or which overlap with multiple clusters, are likely to be outliers on the boundaries of multiple clusters. Figure 4 displays distance ratio distributions corresponding to the settings considered in Table 1. For each algorithm and dataset we show distributions for three subsets of points; those which are uniquely covered by a single cluster, those which are overlapping, and those which are left uncovered. We observe an apparent trend in which the distribution of overlapping and uncovered points is skewed toward smaller distance ratios, indicating they are more likely to be boundary points. This is especially noticeable for `Forest-Depth-4` and the algorithm of `Bandyapadhyay et al.`

Table 1: **Interpretability Measurements** for a setting with 80% required coverage (aside from *Mnist* which is only given a requirement of 60% since it fails to consistently cover much larger). Measurements are computed for maximum rule length (max), weighted average rule length (weighted), coverage, overlap, and partial clustering cost (cost).

Dataset	Method	max	weighted	coverage	overlap	cost
Climate	IMM	5	3.65	1.00	1.00	12.30
	Bandyapadhyay et al.	5	3.58	0.90	1.00	11.49
	IMM-outliers	5	3.61	0.80	1.00	10.67
	Forest-Depth- 4	4	3.37	0.85	1.00	11.35
	Forest-Depth- 3	3	3.00	0.95	1.03	11.60
	Forest-Depth- 2	2	2.00	0.83	1.04	10.94
Anuran	IMM	6	5.32	1.00	1.00	8.73
	Bandyapadhyay et al.	6	5.31	0.86	1.00	6.29
	IMM-outliers	7	5.68	0.80	1.00	5.79
	Forest-Depth- 4	4	4.00	0.80	1.00	5.64
	Forest-Depth- 3	3	3.00	0.81	1.02	6.54
	Forest-Depth- 2	2	2.00	0.81	1.07	7.87
Cover	IMM	4	3.28	1.00	1.00	5.57
	Bandyapadhyay et al.	4	3.29	0.78	1.00	4.58
	IMM-outliers	4	3.28	0.80	1.00	5.10
	Forest-Depth- 4	4	4.00	0.83	1.02	5.04
	Forest-Depth- 3	3	3.00	0.87	1.14	5.50
	Forest-Depth- 2	2	2.00	0.97	1.39	6.23
Digits	IMM	9	6.92	1.00	1.00	3.47
	Bandyapadhyay et al.	9	6.74	0.65	1.00	2.39
	IMM-outliers	8	6.05	0.80	1.00	3.17
	Forest-Depth- 4	4	4.00	0.80	1.05	2.43
	Forest-Depth- 3	3	3.00	0.81	1.12	2.61
	Forest-Depth- 2	2	2.00	0.82	1.26	2.85
Mnist	IMM	8	6.21	1.00	1.00	43.89
	Bandyapadhyay et al.	8	6.13	0.57	1.00	35.02
	IMM-outliers	9	5.78	0.60	1.00	41.64
	Forest-Depth- 4	4	4.00	0.60	1.05	37.57
	Forest-Depth- 3	3	3.00	0.66	1.14	39.15
	Forest-Depth- 2	2	2.00	0.64	1.25	47.37
Fashion	IMM	7	5.05	1.00	1.00	39.30
	Bandyapadhyay et al.	7	5.12	0.71	1.00	29.86
	IMM-outliers	6	4.95	0.80	1.00	36.51
	Forest-Depth- 4	4	3.63	0.83	1.08	32.33
	Forest-Depth- 3	3	3.00	0.81	1.11	33.22
	Forest-Depth- 2	2	2.00	0.82	1.16	35.51

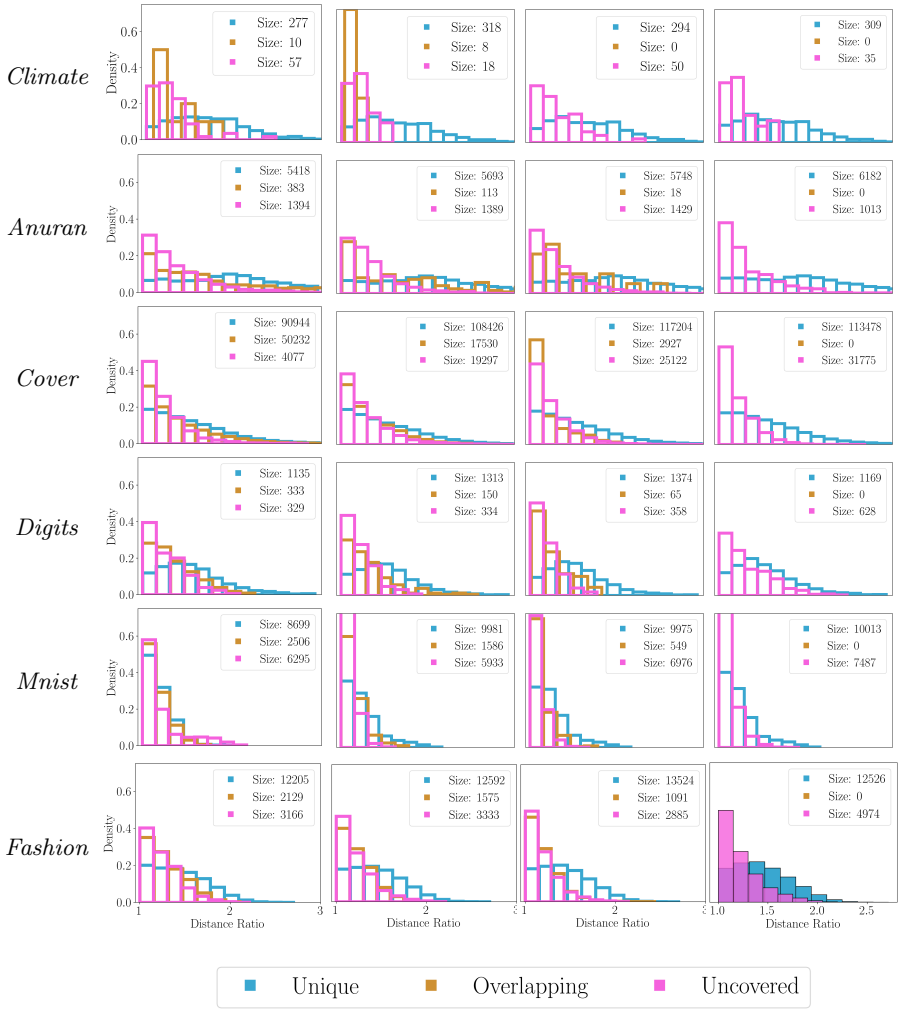


Fig. 4: **Distance Ratio Distributions** for three classes of data points: unique (single cluster), overlapping (multiple clusters), and uncovered (not clustered). Legends in each plot show the respective sizes for each class. Columns correspond to algorithms `Forest-Depth-2-4` and `Bandyapadhyay et al.` respectively. Each row shows results for a given dataset, and the coverage requirements are taken to correspond exactly to those of Table 1.

6 Conclusions

With both qualitative evidence for the *Climate* dataset and experimental results for measurements of distortion and rule length, we have shown that partial clusterings with decision sets are a competitive approach to interpretable clustering problems. Furthermore, we find that random forest generation paired with greedy rule selection is a practical and efficient method for producing low-cost clusterings with descriptions that offer improved user readability in terms of rule length, as compared to previous methods. Finally, while our methods perform well in experimental settings, we argue that the PIC problem remains open to theoretical analysis, and notice that similar problems have been explored in the field of social choice, where limited ordinal information models are compared by cost ratios to their optimal counterparts [3]. An interesting and relevant direction for future study could focus on finding improved solutions or upper bounds in which distortion is computed with respect to an optimal clustering, rather than one produced by an existing algorithm.

References

1. Bandyapadhyay, S., Fomin, F.V., Golovach, P.A., Lochet, W., Purohit, N., Simonov, K.: How to find a good explanation for clustering? Artificial Intelligence **322**, 103948 (2023)
2. Blackard, J.: Covertype. UCI Machine Learning Repository (1998), DOI: https://doi.org/10.24432/C50K5N
3. Burkhardt, J., Caragiannis, I., Fehrs, K., Russo, M., Schwiegelshohn, C., Shyam, S.: Low-distortion clustering with ordinal and limited cardinal information. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 9555–9563 (2024)
4. Carrizosa, E., Kurishchenko, K., Marín, A., Morales, D.R.: On clustering and interpreting with rules by means of mathematical optimization. Computers & Operations Research **154**, 106180 (2023)
5. Chawla, S., Gionis, A.: k-means–: A unified approach to clustering and outlier detection. In: Proceedings of the 2013 SIAM international conference on data mining. pp. 189–197. SIAM (2013)
6. Chen, J., Chang, Y., Hobbs, B., Castaldi, P., Cho, M., Silverman, E., Dy, J.: Interpretable clustering via discriminative rectangle mixture model. In: 2016 IEEE 16th international conference on data mining (ICDM). pp. 823–828. IEEE (2016)
7. Colonna, J., Nakamura, E., Cristo, M., Gordo, M.: Anuran Calls (MFCCs). UCI Machine Learning Repository (2015), DOI: https://doi.org/10.24432/C5CC9H
8. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012)
9. Frost, N., Moshkovitz, M., Rashtchian, C.: Exkmc: Expanding explainable k-means clustering. arXiv preprint arXiv:2006.02399 (2020)
10. Gabidolla, M., Carreira-Perpiñán, M.Á.: Optimal interpretable clustering using oblique decision trees. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 400–410 (2022)
11. Gamlath, B., Jia, X., Polak, A., Svensson, O.: Nearly-tight and oblivious algorithms for explainable clustering. Advances in Neural Information Processing Systems **34**, 28929–28939 (2021)

12. Gu, S., Chou, Y., Zhou, J., Jiang, Z., Lu, M.: Takagi–sugeno–kang fuzzy clustering by direct fuzzy inference on fuzzy rules. IEEE Transactions on Emerging Topics in Computational Intelligence **8**(2), 1264–1279 (2023)
13. Guilbert, M., Vrain, C., Dao, T.B.H.: Towards explainable clustering: A constrained declarative based approach. arXiv preprint arXiv:2403.18101 (2024)
14. Harshaw, C., Feldman, M., Ward, J., Karbasi, A.: Submodular maximization beyond non-negativity: Guarantees, fast algorithms, and applications. In: International Conference on Machine Learning. pp. 2634–2643. PMLR (2019)
15. Hu, L., Jiang, M., Dong, J., Liu, X., He, Z.: Interpretable clustering: A survey. arXiv preprint arXiv:2409.00743 (2024)
16. Hwang, H., Whang, S.E.: Xclusters: explainability-first clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 7962–7970 (2023)
17. Laber, E., Murtinho, L., Oliveira, F.: Shallow decision trees for explainable k-means clustering. Pattern Recognition **137**, 109239 (2023)
18. Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets: A joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1675–1684 (2016)
19. Lawless, C., Gunluk, O.: Cluster explanation via polyhedral descriptions. In: International conference on machine learning. pp. 18652–18666. PMLR (2023)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
21. Liu, H., Li, J., Wu, Y., Fu, Y.: Clustering with outlier removal. IEEE transactions on knowledge and data engineering **33**(6), 2369–2379 (2019)
22. Mansoori, E.G.: Frbc: A fuzzy rule-based clustering algorithm. IEEE transactions on fuzzy systems **19**(5), 960–971 (2011)
23. Moshkovitz, M., Dasgupta, S., Rashtchian, C., Frost, N.: Explainable k-means and k-medians clustering. In: International conference on machine learning. pp. 7055–7065. PMLR (2020)
24. National Oceanic Atmospheric Administration (NOAA): Climate at a Glance: U.S. Regional Time Series. https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/divisional/mapping (2025), accessed: 2025-03-14
25. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
26. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys **16**, 1–85 (2022)
27. Saisubramanian, S., Galhotra, S., Zilberstein, S.: Balancing the tradeoff between clustering value and interpretability. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 351–357 (2020)
28. Sathiaraj, D., Huang, X., Chen, J.: Predicting climate types for the continental united states using unsupervised clustering techniques. Environmetrics **30**(4), e2524 (2019)
29. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
30. Zeng, T., Zhong, C., Pan, T.: Clustering explanation based on multi-hyperrectangle. Scientific Reports **14**(1), 1–18 (2024)
31. Zhang, G., Gionis, A.: Diverse rule sets. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1532–1541 (2020)