EXPLAINABLE EVIDENTIAL CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised classification is a core problem in machine learning. Because real-world data are often imperfect, non-additive frameworks, such as evidential clustering, grounded in Dempster-Shafer theory, explicitly handle uncertainty and imprecision. These frameworks are particularly well suited to high-stakes decisions, which tend to require both interpretability and cautiousness. However, while decision-tree surrogates have enabled transparent explanations for hard clustering, explainability for evidential clustering remains largely unexplored. We address this gap by formalizing representativeness, a utility-based criterion that captures decision-makers' preferences over explanation misassignments, and introducing evidential mistakeness, a loss function tailored to credal partitions. Building on these foundations, we propose the Iterative Evidential Mistakeness Minimization (IEMM) algorithm, which learns decision-tree explainers for evidential clustering by optimizing representativeness under uncertainty and imprecision. We provide theoretical conditions for effective explanations in both hard and evidential settings and show how utility function parameters can be set to reflect different decision attitudes. Experiments on synthetic and real-world datasets demonstrate that IEMM improves the performance of existing methods by producing representative and preference-aligned explanations of evidential clusterings, supporting cautious, transparent analysis in the presence of imperfect data.

1 INTRODUCTION

Clustering is a fundamental machine learning problem MacQueen (1967) that aims to group similar objects while distinguishing different ones Hansen & Jaumard (1997). As a core data analysis task, it reveals patterns and enables applications such as data compression, summarization, visualization, and anomaly detection Xu & Wunsch (2005). As with other machine learning methods, two major challenges persist in clustering: **imperfections in the input data** Hüllermeier & Waegeman (2021) and **interpretability** Carvalho et al. (2019).

Real-world scenarios with imperfect data require **cautiousness** Bengs et al. (2022); Angelopoulos et al. (2022); Imoussaten & Jacquin (2022); Hüllermeier et al. (2022); Nguyen et al. (2018), defined as decision-makers' awareness of model limitations and resulting risk-aversion. Effective cautiousness depends on properly characterizing these imperfections, primarily **uncertainty** and **imprecision** Dubois & Prade (2009). In machine learning contexts, imperfections typically arise from weak supervision, aleatoric uncertainty (intrinsic variability in the data), and epistemic uncertainty (a lack of data in parts of the feature space) Hüllermeier & Waegeman (2021). Approaches that address these issues include imprecise probability theory Walley (1991), possibility theory Dubois & Prade (1988), rough sets Pawlak (1982), fuzzy sets Zadeh (1965), and Dempster-Shafer evidence theory Shafer (1976).

These foundations have given rise to various clustering methods, including fuzzy Ruspini (1969), possibilistic Krishnapuram & Keller (1993), rough Lingras & West (2004), and evidential clustering Masson & Denœux (2008). In the latter framework, while classical hard clustering Hartigan & Wong (1979) assigns each point to exactly one cluster, evidential clustering induces a credal partition Masson & Denœux (2008) that represents both uncertainty and imprecision through partial membership across multiple cluster combinations.

Interpreting clustering results is equally critical: without interpretation, clusterings often lack practical utility. This has motivated a growing body of work on interpretable clustering Ben-Hur et al.

(2001); Carrizosa et al. (2022); Lawless & Gunluk (2022); Ellis et al. (2021; 2024); Alvarez-Garcia et al. (2024); Tutay & Somech (2023). A prominent approach borrows from supervised learning: treat the (hard) clustering labels as ground truth and train a surrogate decision-tree classifier to reproduce them. A notable example is Iterative Mistake Minimization (IMM) Moshkovitz et al. (2020), which fits a decision tree aligned with the centroid structure—each leaf holds exactly one centroid, mapping its points to the associated cluster. Building on the IMM, subsequent work strengthens theoretical guarantees of the provided explanation and broadens its scope Makarychev & Shan (2022); Frost et al. (2020); Laber et al. (2022); Bandyapadhyay et al. (2023); Gabidolla & Carreira-Perpiñán (2022); Fleissner et al. (2024).

Explainability refers to a model's ability to provide clear, audience-appropriate reasons for its behavior Barredo Arrieta et al. (2020). It enables users to understand, critique, and improve models. A common taxonomy distinguishes **intrinsic** methods—models designed to be interpretable—from **post-hoc** methods—explanations for already trained black boxes Carvalho et al. (2019). Post-hoc techniques mainly fall into two families Barredo Arrieta et al. (2020): (i) feature-relevance methods, which rank or quantify the influence of input features Lundberg et al. (2019); Baehrens et al. (2010) but, because they reveal little about the dataset's structure Moshkovitz et al. (2020), face criticism in high-stakes settings Rudin (2019); and (ii) **simplification** methods, which approximate black-box classifiers with interpretable surrogates, such as decision trees, rule lists, or linear models Guidotti et al. (2018). These categories are not mutually exclusive: explanations may also be delivered through examples, counterfactuals, or visual/textual modalities Barredo Arrieta et al. (2020). Nor are they exhaustive. For example, feature relevance can be obtained via simplification, as in LIME Ribeiro et al. (2016). Our proposed approach can be viewed as both intrinsic (it produces interpretable models) and post-hoc simplification (it explains a given clustering).

High-stakes domains such as healthcare demand both interpretability and cautiousness. Yet explaining cautious clustering remains largely unexplored. Only a few works extend explainability to imprecise methods, with early efforts focusing on supervised classification via counterfactuals and feature importance Zhang (2023). To our knowledge, no prior work explains evidential clustering. As noted in literature Zhang et al. (2024), explainable clustering over uncertain or imprecise data warrants investigation to enable cautious, transparent analysis under imperfect data sources. This paper addresses that gap.

The main **objectives** of this paper are:

1. To conduct a comprehensive investigation of decision trees as explainers for hard clustering functions, establishing conditions that define effective explanations.
2. To develop a theoretical framework that extends these conditions to encompass uncertainty and imprecision, particularly within the evidential clustering paradigm.
3. To introduce an innovative **Explainable Evidential Clustering** method through a novel algorithm grounded in these theoretical foundations.

Our key **contributions** include:

1. We demonstrate that representativity is a necessary and sufficient condition for decision trees to act as abductive explainers in the hard case. Building upon utility functions, we introduce the concept of *Evidential Representativeness*, which quantifies decision-makers' preferences regarding errors committed by an explainer. This advancement enables systematic evaluation of cautious explanations.
2. We propose the Evidential Mistakeness function, demonstrate that minimizing it leads to representative explanations, and develop the **Iterative Evidential Mistakeness Minimization** (IEMM) algorithm. This novel approach, inspired by the IMM algorithm Moshkovitz et al. (2020), generates surrogate decision trees that effectively explain evidential clustering functions.
3. We implement and validate this algorithm on both synthetic and real-world datasets, demonstrating how to select utility parameters that reflect different decision attitudes.

The remainder of this paper is structured as follows. Section 2 presents the theoretical foundations, encompassing belief functions, evidential clustering, and explainability. Section 3 introduces the concepts on which our approach is based: a specific family of utility functions, the representative-

ness criterion, the Evidential Mistakeness loss function, and the IEMM algorithm. We provide formal analysis of these concepts along with illustrative examples. We also make available the IEMM Python package and the complete code for all experiments at OMITTED TO AVOID IDENTIFICATION.

2 BACKGROUND

Let X represent a set of **observations** in a known **feature space** $\mathbb{X}$. We assume $X = \{x_1, ..., x_N\} \subset \mathbb{X} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_D$, where each element of $\mathcal{D} = \{\mathcal{A}_1, \ldots, \mathcal{A}_D\}$ is called an **attribute**. We assume all attributes are finite[1]. In essence, X is a set of D measurements for each of N objects, while $\mathbb{X}$ encompasses all possible measurements.

A classification problem is the task of assigning each observation in X to an outcome from a finite set $\Omega = \{\omega_1, ..., \omega_C\}$, which we call the **frame of discernment**. A function that performs this assignment is called a **classifier**. When a set of training examples is available, we refer to the problem of constructing such function as **supervised classification**. In contrast, if no training examples are available and the goal is to group observations based on their similarity—without prior knowledge of the classes or labels—the task is called **unsupervised classification** or **clustering**.

2.1 BELIEF FUNCTIONS

The Dempster-Shafer theory of evidence Shafer (1976) provides a framework for representing uncertain and imprecise information. At the core of this theory lies the **mass of belief function**, or simply **mass function**—a map defined as:

$$m : 2^\Omega \to [0, 1] \text{ such that } \sum_{A \subseteq \Omega} m(A) = 1.$$

Within this framework, an element $\omega \in \Omega$ represents the finest level of discernible information. The mass $m(A)$ quantifies the degree of confidence in the statement that 'the correct hypothesis ω belongs to $A \subseteq \Omega$, yet it remains impossible to determine which specific element of A is correct'. When $m(\varnothing) = 0$, we say the mass *satisfies the closed-world hypothesis* Smets (1988), meaning it rejects the possibility that the correct hypothesis ω lies outside Ω.

We define the **focal set** of m as $\mathbb{F}_m = m^{-1}(]0, 1])$—the collection of all subsets of Ω assigned nonzero belief. Each member of this set is known as a **focal element**. Mass functions can be categorized based on their focal elements:

- If all focal elements are singletons (of cardinality 1), then $p(\omega) = m(\{\omega\})$ forms a probability mass function, and m is called a Bayesian mass function.
- A mass function with exactly one focal element A is called *categorical*, representing the logical assertion that 'ω belongs to A'. If this single focal element is Ω itself, the function is *vacuous*, conveying no information beyond the closed-world hypothesis.

We denote by $\mathbb{M}$ the set of all mass functions defined on Ω. For notational simplicity, we may write $\omega_i \cup \omega_j \cup \omega_k \cup ...$ to represent the subset $\{\omega_i, \omega_j, \omega_k, ...\}$. In Appendix A we present some additional constructions (belief/plausability/pignistic functions) allowing decision-making for a mass of belief function.

2.2 EVIDENTIAL CLUSTERING

Definition 1. An **evidential clustering** is a map $\mathcal{M} : X \to \mathbb{M}$.

For an observation $x \in X$, the function $\mathcal{M}(x)$, which we may denote as m_x, when evaluated at $A \subseteq \Omega$, returns the degree of confidence attributed to the statement 'the class ω corresponding to

[1]This finite attribute assumption is crucial for decision tree operations. This assumption reflects this work's aim of explaining the clustering of tabular data. We may occasionally refer to $\mathbb{X}$ as $\mathbb{R}^D$ for simplicity, though this is an abuse of notation. When working with continuous attributes, we implicitly discretize the space (for example, with a dataset $X \subset \mathbb{R}^D$, we typically consider binary attributes $\mathcal{A}_{d,\theta} = \{True, False\}$ for each dimension $d \in \{1, ..., D\}$ and threshold $\theta \in \{x_d : x \in X\}$, where $x_{\mathcal{A}_{d,\theta}} = True$ if and only if $x_d \geq \theta$).

x belongs to A, and it is not possible, given the available information, to determine which specific element of A is the correct one'.

We refer to each element of Ω as a cluster and each subset of Ω as a metacluster. Within the context of an evidential clustering function, we denote $\mathbb{F}_{\mathcal{M}} = \bigcup_{x \in X} \mathbb{F}_{m_x}$. When all m_x are categorical, we say that $\mathcal{M}$ is *categorical*. Similarly, when all elements of $\mathbb{F}_{\mathcal{M}}$ are singletons, we call $\mathcal{M}$ *bayesian*. An evidential clustering function that is both categorical and bayesian naturally induces a hard clustering. A **hard clustering** is simply a partition of observations into clusters, formalized as a surjection $\mathcal{C} : X \to \Omega$. Appendix A further discusses evidential clustering and offers a visualization. Some clustering algorithms naturally produce a centroid for each cluster. A **centroid** v_ω (resp. v_A) is a point in the feature space $\mathbb{X}$ that represents its cluster $\omega \in \Omega$ (resp. metacluster $A \subset \Omega$). For the remainder of this work, we assume $\varnothing \notin \mathbb{F}_{\mathcal{M}}$, rejecting the outlier hypothesis.

2.3 Decision Trees as Explainers

Decision Trees (DTs) Quinlan (1987) are classical machine learning algorithms, classifiers based on rooted computation trees expressed as recursive partitions of the observation space Rokach & Maimon (2005). Building upon these partitional aspects, a **node** of a decision tree is defined as the subset $S \subseteq \mathbb{X}$ associated with its vertices. Decision trees are widely used in explainability for their inherent interpretability Barredo Arrieta et al. (2020), as they yield "a set of decision rules with the if–then form" Guidotti et al. (2018). In this context, a particularly desirable outcome Amgoud & Ben-Naim (2022) is an abductive explanation: it answers the question "Why is $\Gamma(x) = \omega$?" by providing a sufficient reason for assigning the label ω, where $\Gamma : \mathbb{X} \to \Omega$ is a supervised classifier Ignatiev et al. (2019). Throughout the paper, we denote by $\mathbf{C}$ the set of all consistent subsets of feature literals[2] and call an explainer any map $\chi_\Gamma : \Omega \to 2^{\mathbf{C}}$. Further definitions and formal aspects of explainers are provided in Appendix B.

A decision tree $\Delta : \mathbb{X} \to \Omega$ induces an explainer $\chi_\Gamma^\Delta : \Omega \to 2^{\mathbf{C}}$ that provides abductive explanations to a supervised classifier Γ if and only if $\Delta = \Gamma$. This equality between the original classifier and the DT surrogate model is equivalent to the property that Amgoud & Ben-Naim (2022) calls representativity. A representative explainer is one that, for all observations x with label $\omega = \Gamma(x)$, can provide an explanation in $\chi_\Gamma^\Delta(\omega)$ that holds at x. More details on the DT construction from the standpoint of explainability can be found in appendix C, along with our proposed proof that representativity is a necessary and sufficient condition for decision trees to provide abductive explanations.

As the complete representativity is rarely attainable, it is natural to assess the quality of explanations by the "representativeness" of the explainer χ_Γ^Δ. This assessment, in the supervised case, is typically performed by measuring the accuracy Ribeiro et al. (2016); Izza et al. (2022b); Narodytska et al. (2019) of the underlying classifier Δ. Thus, the quality of the explanation provided by χ_Γ^Δ about Γ is quantified as:

$$\text{Accuracy}_\Gamma(\Delta) = \frac{|\{x \in X : \Gamma(x) = \Delta(x)\}|}{|X|}. \tag{1}$$

2.3.1 The IMM: Decision Trees Explaining Hard Clustering

Iterative Mistake Minimization (IMM) Moshkovitz et al. (2020) explains a hard clustering by training a decision tree that mimics the cluster assignments (see Figure 1) while minimizing the price of explanation. To this end, it relies on the concept of *mistake*.

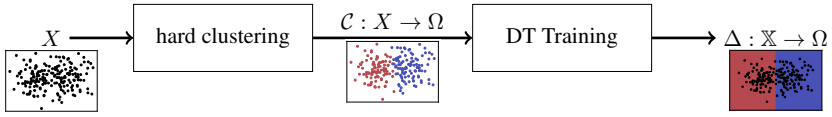

Figure 1: Schematic of explainable clustering.

Definition 2. Let $\mathcal{C}$ be a hard clustering function. A **mistake** in a decision-tree (DT) node $S \subseteq \mathbb{X}$ occurs when a point $x \in S$ has its associated cluster centroid $v_{\mathcal{C}(x)}$ outside of S, i.e., $v_{\mathcal{C}(x)} \notin S$.

[2] A feature literal is a pair $\langle \mathcal{A}, v \rangle$ where $\mathcal{A} \in \mathcal{D}$ and $v \in \mathcal{A}$. A consistent subset of feature literals is some set L of feature literals such that $\langle \mathcal{A}, v \rangle, \langle \mathcal{A}, v' \rangle \in L \Rightarrow v = v'$.

The **number of mistakes** in a decision tree is the sum of mistakes committed with respect to $\mathcal{C}$ across all leaves. The IMM aims to greedily minimize the number of mistakes induced by each axis-aligned split in the decision tree Moshkovitz et al. (2020).

To assess explanation quality, as discussed in the previous section, it is natural to measure how representative the resulting explainer is. The original IMM evaluation Moshkovitz et al. (2020) relied on an explanation cost derived from the k-means and k-medians objectives. Some works Fleissner et al. (2024); Lawless & Gunluk (2022)—not restricted to these clustering methods—use the Rand index Rand (1971) or the $\text{Accuracy}_{\mathcal{C}}(\Delta)$ to assess the similarity between the original clustering and the clustering induced by the decision tree.

3 An Algorithm for Explaining Evidential Clustering

Our objective is to extend the concept of decision trees for cluster explanation to the evidential setting. We aim to construct a decision tree that provides a representative approximation of the evidential clustering function, as illustrated in Figure 2.

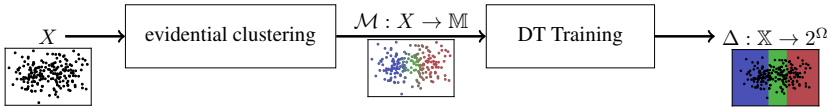

Figure 2: Scheme of Explainable Evidential Clustering.

To achieve this, we build on the hard case: we first generalize what it means for an explainer to be representative, and then, inspired by IMM, derive the notion of a mistake for evidential classifiers and propose an algorithm that seeks to minimize the resulting loss. We start with the specific case of categorical mass functions and then extend to the general case.

3.1 Explaining Classifiers under Uncertainty and Imprecision

Given an evidential classifier[3] $\mathcal{M} : X \to \mathbb{M}$, we seek to construct an interpretable decision tree $\Delta : \mathbb{X} \to 2^{\Omega}$ that approximates $\mathcal{M}$. We first formalize this approximation based on the quality of the generated explanations.

3.1.1 Utility Functions

Let us consider a categorical evidential classifier $\mathcal{M}_c : X \to \mathbb{M}$. Let us define $\overline{\mathcal{M}}_c : X \to 2^{\Omega}$ such that, for any $x \in X$ and $A \subseteq \Omega$, $\mathcal{M}_c(x)(A) = 1$ if and only if $\overline{\mathcal{M}}_c(x) = A$.

Assessing the quality of a surrogate partition Δ is challenging in the evidential setting: predictions and truths are subsets of Ω, so errors vary in severity. For example, predicting $\{\omega_1, \omega_2\}$ when $\overline{\mathcal{M}}_c(x) = \{\omega_1\}$ is arguably less severe than predicting $\{\omega_2\}$, yet exact-match representativeness treats both equally. Such binary criteria ignore partial agreement and domain-specific preferences across clusters. To make these trade-offs explicit, we introduce a specific family of bounded utility functions that quantifies the decision-maker's satisfaction when A is predicted while the ground truth is B.

Definition 3. A **utility function** is a map $\mathcal{U} : 2^{\Omega} \times 2^{\Omega} \to [0, 1]$ such that, $\forall A, B \in 2^{\Omega}$,

a) $\mathcal{U}(A, A) = 1$ and
b) $A \cap B = \varnothing \Rightarrow \mathcal{U}(A, B) = 0$.

Utility functions are standard tools in decision theory Keeney & Raiffa (1993) and the existence of utility functions that encode decision-maker's (DM's) preferences has been widely discussed Von Neumann & Morgenstern (1947); Savage (1954); Schmeidler (1989). In the evidential setting, several works explore desirable properties and elicitation procedures for utilities. Some of them focused on the case where ground truth is assumed to be known precisely Zaffalon et al. (2012); Ma & Denœux (2021) and others Jacquin et al. (2019); Imoussaten & Jacquin (2022); Imoussaten

[3]We use the term classifier to emphasize that the development of this section is valid not only for clustering but also for all evidential partitions of the data.

(2023) extend to the cases where ground truth is itself imprecise. In our formulation, we consider a bounded utility as described by conditions a) and b) in Definition 3: a perfect prediction yields total satisfaction for the DM (utility equals 1), whereas completely disjoint prediction and truth yield unacceptable satisfaction for the DM (utility equals 0). This assumption is convenient for our analysis but is not the general case Kunitomo-Jacquin et al. (2025). In Appendix D, we discuss how such utilities relate to explanation costs of misassignments in the categorical case and provide an illustrative example.

3.1.2 Evidential Representativeness

Definition 4. A **cautious explainer** for a categorical evidential classifier $\mathcal{M}_c : X \to \mathbb{M}$ is a map $\chi_{\mathcal{M}_c} : 2^\Omega \to 2^{\mathbf{C}}$.

The cautious explainer differs from a standard explainer in its capability to explain the classifier's imprecise predictions. In this context, a good explainer should ensure that if x is assigned to A by the classifier, then there exists some B such that $\mathcal{U}(A, B) = 1$ and x is explained by $\chi_{\mathcal{M}_c}(B)$. We therefore characterize the representativeness of a cautious explainer with respect to a utility.

Definition 5. A **$\mathcal{U}$-representative cautious explainer** for a categorical evidential clustering is a cautious explainer $\chi_{\mathcal{M}_c}$ such that, $\forall A \in 2^\Omega$, $\forall x \in \overline{\mathcal{M}}_c^{-1}(\{A\})$, there exists $L \in \bigcup_{\mathcal{U}(B,A)=1} \chi_{\mathcal{M}_c}(B)$ such that, $\forall \langle \mathcal{A}, v \rangle \in L$, $x_{\mathcal{A}} = v$.

Different utility functions induce different notions of representativeness. Intuitively, more permissive utilities yield higher scores and tolerate more error. In the categorical evidential case, utility allows us to define the **$\mathcal{U}$-categorical representativeness** of a cautious explainer as its average utility:

$$\mathcal{R}_{\mathcal{M}_c,\mathcal{U}}(\Delta) = \frac{1}{|X|} \sum_{x \in X} \mathcal{U}(\Delta(x), \overline{\mathcal{M}}_c(x)). \tag{2}$$

For any evidential partition $\mathcal{M} : X \to \mathbb{M}$, we can extend Equation 2. Let us define then the **$\mathcal{U}$-evidential representativeness** $\mathcal{R}_{\mathcal{M},\mathcal{U}} : (\mathbb{X} \to 2^\Omega) \to [0, 1]$ of a cautious explainer as the expected categorical representativeness weighted by the mass function:

$$\mathcal{R}_{\mathcal{M},\mathcal{U}}(\Delta) = \frac{1}{|X|} \sum_{x \in X} \sum_{B \in \mathbb{F}_{\mathcal{M}}} \mathcal{U}(\Delta(x), B) m_x(B). \tag{3}$$

Equation (2) is clearly a special case of (3) when $\mathcal{M}$ is categorical. Additionally, equations (3) and (1) coincide when $\mathcal{M}$ is a hard partition and $\mathcal{U}(A, B) = \mathbb{1}_{A=B}$.

The literature offers several ways to compare evidential partitions, mainly via distances between mass functions or by aggregating nonspecificity with measures of conflict Jousselme et al. (2001); Jousselme & Maupin (2012); Hoarau et al. (2023a); Denoux et al. (2018); Campagner et al. (2023); Masson & Denœux (2008). However, we believe that utility-based representativeness offers advantages for our setting. As we face an explanation task, the choice of utilities lets us encode decision-maker preferences in a more interpretable way than tuning parameters of a clustering objective Masson & Denœux (2008) or weighting nonspecificity and conflict Denoux et al. (2018); Denoeux & Bjanger (2000), as those values often lack an immediate meaning to the explanation audience. Moreover, unlike distances, utilities can capture complex and possibly asymmetric preferences. For instance, mapping $A = \{\omega_1\}$ to $B = \{\omega_1, \omega_2\}$ need not be penalized the same as mapping $A = \{\omega_1, \omega_2\}$ to $B = \{\omega_1\}$.

3.1.3 Evidential Mistakeness

With this updated notion of representativeness, we can extend the concept of a mistake to the evidential case as a cost function capturing the representativeness loss associated with a single DT explainer node. Recall from Definition 2 that, in the hard case, the number of mistakes in a node S can be described in two equivalent ways:

1. The number of mistakes in S is the number of points $x \in S$ such that $\exists v_\omega \notin S$ with $\omega = \mathcal{C}(x)$.
2. The number of mistakes in S is the number of points $x \in S$ such that $\forall v_\omega \in S, \omega \neq \mathcal{C}(x)$.

Translating these to the evidential setting yields two natural definitions of **evidential mistakeness**:

1. The evidential mistakeness in S is the sum of the costs introduced by not assigning points x in S to metaclusters that are not in S:
$$\overline{M}_{\mathcal{M},\mathcal{U}}(S) = \sum_{x \in S} \sum_{v_A \notin S} \sum_{B \in \mathbb{F}_{\mathcal{M}}} \mathcal{U}(A, B) m_x(B). \tag{4}$$
2. The evidential mistakeness in S is the sum, over all x in S, of the expected cost of assigning x to some metacluster in S:
$$\underline{M}_{\mathcal{M},\mathcal{U}}(S) = \sum_{x \in S} \sum_{v_A \in S} \sum_{B \in \mathbb{F}_{\mathcal{M}}} \frac{(1 - \mathcal{U}(A, B)) m_x(B)}{|\{C \in \mathbb{F}_{\mathcal{M}} : v_C \in S\}|} \tag{5}$$

One can interpret Equation (4) as the satisfaction that the DM will not concretize and Equation (5) as the unsatisfaction that the DM will concretize. When $\mathcal{M}$ induces a hard clustering and $\mathcal{U}(A, B) = \mathbb{1}_{A=B}$, Equation (4) equals the number of mistakes in S. Additionally, the total cost of a cautious DT explainer induced by Equation (5) is zero if and only if the explainer is $\mathcal{U}$-representative.

For IMM-like algorithms where all leaves S contain exactly one centroid, both evidential mistakeness forms from Equations (4) and (5) are minimized by explanations with maximal evidential representativeness (see proof in Appendix E). The key difference between these definitions emerges in nodes containing multiple centroids, where Equation (5) penalizes such nodes more heavily than Equation (4). This makes Equation (4) better suited for conservative explainers, while Equation (5) is preferable for more risk-attractive ones.

3.2 The Algorithm

Algorithm 1 IEMM

Input: Observations $X = [x^1, \ldots, x^N] \subset \mathbb{R}^D$.
Some evidential clustering $\mathcal{M} : X \to \mathbb{M}$.
The focal sets $\mathbb{F} = \{A_1, \ldots, A_{|\mathbb{F}|}\}$
and their centroids $v = \{v^1, \ldots, v^{|\mathbb{F}|}\} \subset \mathbb{R}^D$.
Parameter: The chosen evidential mistakeness **M**
Output: A decision tree $\Delta : \mathbb{R}^D \to 2^\Omega$.

$\Delta \leftarrow$ **split_tree**$(X, \mathcal{M}(X), \mathbb{F}, v)$
function SPLIT_TREE($\{x^j\}_{j=1}^n, \{m^j\}_{j=1}^n, F, \{v^j\}_{A_j \in F}$)
  if $|F| = 1$ **then**
    leaf.metacluster $\leftarrow F$
    return leaf
  end if
  for all $i \in [1, \ldots, D]$ **do**
    $\ell_i \leftarrow \min_{A_j \in F} v_i^j$
    $r_i \leftarrow \max_{A_j \in F} v_i^j$
  end for
  $i, \theta \leftarrow \arg\min_{i, \ell_i \le \theta < r_i} \mathbf{M}(x, m, v, F, i, \theta)$
  $L \leftarrow \{j \mid (x_i^j \le \theta)\}_{j=1}^n$
  $R \leftarrow \{j \mid (x_i^j > \theta)\}_{j=1}^n$
  $F_L \leftarrow \{A_j \in F \mid v_i^j \le \theta\}$
  $F_R \leftarrow \{A_j \in F \mid v_i^j > \theta\}$
  node.condition $\leftarrow$ "$x_i \le \theta$"
  node.lt $\leftarrow$ **split_tree**$(\{x^j\}_{j \in L}, \{m^j\}_{j \in L}, F_L, v)$
  node.rt $\leftarrow$ **split_tree**$(\{x^j\}_{j \in R}, \{m^j\}_{j \in R}, F_R, v)$
  return node
end function

Inspired by IMM, we propose the **Iterative Evidential Mistakeness Minimization (IEMM)**. The IEMM fits a decision tree based on an evidential clustering by minimizing the evidential mistakeness function (see Algorithm 1). Each iteration of IEMM, for a region $S \subseteq \mathbb{X}$, considers a subset $F \subseteq \mathbb{F}_{\mathcal{M}}$ of the focal sets whose metacluster centroids lie within S and finds the split that, by separating centroids, minimizes the contribution to the evidential mistakeness.

From a computational perspective, the baseline IMM algorithm has complexity $O(C \cdot D \cdot N \cdot \log N)$ Moshkovitz et al. (2020), where $C = |\Omega|$ represents the number of clusters, D the dimensionality of the feature space, and N the sample count. Our IEMM algorithm extends this by incorporating utility computations at each node, adding an $O(K^2)$ factor where $K = |\mathbb{F}|$ is the number of metaclusters. This results in a total complexity of $O(K^2 \cdot D \cdot N \cdot \log N)$. In practice, explainable decision trees typically employ a modest number of metaclusters, mitigating potential performance concerns.

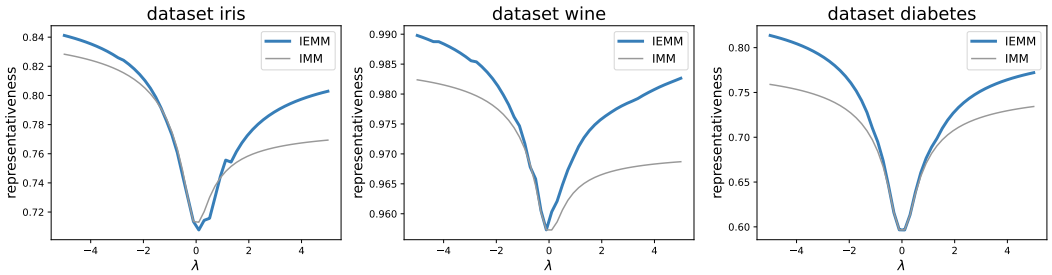

Figure 3: Obtained $\mathcal{U}^\lambda$-evidential representativeness $\mathcal{R}_{\mathcal{M},\mathcal{U}^\lambda}(\Delta)$ when explaining an ECM clustering across multiple datasets. For each dataset, we run ECM Masson & Denœux (2008) to obtain an evidential clustering function $\mathcal{M}$ and compare, for various λ, the Δ_{IEMM} learned under mistakeness $M^\lambda_{\mathcal{M}}$ with Δ_{IMM}, produced by applying the adapted baseline IMM Moshkovitz et al. (2020).

3.2.1 Experiments

We evaluate whether IEMM produces explanations that align with a decision-maker's (DM's) preference for cautiousness. Preferences are modeled via a utility function family $\{\mathcal{U}^\lambda\}_{\lambda\in\mathbb{R}\cup\{\pm\infty\}}$, introduced in Appendix F, which captures different risk attitudes: larger λ values correspond to more cautious choices, trading conflict for non-specificity in the spirit of Denoux et al. (2018). In practical applications, the utility should be elicited directly with the DM Kunitomo-Jacquin et al. (2025); here we vary λ to study behavior across a spectrum of attitudes.

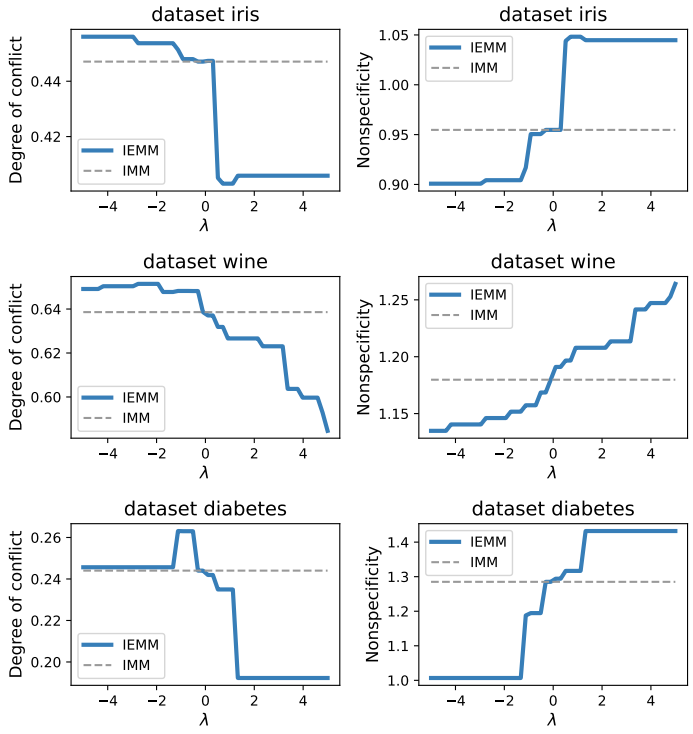

Figure 4: Conflict/non-specificity analysis for IEMM and IMM across λ values. Higher λ yields more cautious (less conflicting but more non-specific) explanations; lower λ emphasizes specificity. Metrics are computed following Denoux et al. (2018).

The λ-evidential mistakeness function $M^\lambda_{\mathcal{M}}$ (as described in Appendix F) is designed to be consistent with this utility family. At $\lambda = 0$, all errors are weighted equally, recovering the original IMM mistakeness in the hard-clustering case.

Because IEMM is, to our knowledge, the first algorithm able to explain cautious partitions, we compare it against a careful adaptation of IMM Moshkovitz et al. (2020). Given an evidential clustering function $\mathcal{M}$, we derive a categorical partition by applying the strong dominance criterion pointwise (Appendix A), and then run IMM while treating each metacluster as an ordinary cluster.

Experimental protocol: For each dataset (Iris Fisher (1936), Wine Aeberhard et al. (1991), and Diabetes Efron et al. (2004) from `sklearn` Pedregosa et al. (2011)), we first run ECM Masson & Denœux (2008) to obtain $\mathcal{M}$. For a grid of λ values, we learn Δ_{IEMM} by minimizing $M^\lambda_{\mathcal{M}}$ and obtain Δ_{IMM} from the induced categorical partition. We then measure $\mathcal{U}^\lambda$-evidential representativeness $\mathcal{R}_{\mathcal{M},\mathcal{U}^\lambda}(\Delta)$ for both explainers. Results are summarized in Figure 3.

Main findings: IEMM consistently achieves higher $\mathcal{U}^\lambda$-evidential representativeness than IMM, with the difference vanishing at $\lambda = 0$. This indicates that IEMM better preserves the cautiousness inherent in the evidential clustering, while remaining competitive when cautiousness is not emphasized. This highlights the IEMM's ability to adapt explanations to the DM's risk attitude, effectively balancing specificity and cautiousness as desired, while IMM lacks this flexibility.

Trade-off analysis: Conflict and non-specificity define antagonistic objectives in evidential clustering Denoux et al. (2018). Varying λ smoothly navigates this frontier: increasing λ promotes cautious (lower-conflict, higher non-specificity) explanations, whereas decreasing λ prioritizes specificity. Figure 4 illustrates how IEMM move along this trade-off as λ changes, while IMM remains a static reference, as it does not adapt to the DM's utility function.

Additional experimental details and further tests on synthetic and real-world datasets are provided in Appendix G.

4 Conclusion

In this paper, we presented a novel approach to explainable evidential clustering using decision trees as explainers. Through the introduction of utility functions, we extended the concept of representativity, a both necessary and sufficient condition for decision trees to function as abductive explainers, to imprecise settings. This allows for the accommodation of "tolerable" mistakes in explanations, making it particularly suitable for evidential contexts. Building on these theoretical foundations, we proposed the evidential mistakeness measure and developed the Iterative Evidential Mistakeness Minimization (IEMM) algorithm. Our approach produces decision trees that effectively explain evidential clustering, advancing the development of both cautious and explainable AI systems.

An important consideration regards the expected audience of the explanations our algorithm creates. Our work implicitly assumes that decision-makers possess domain expertise, an understanding of the implications of their choices, and knowledge about their risk tolerance preferences. The explanations we generate are designed for these informed stakeholders—individuals familiar with the feature space and its relationships. For example, in clinical applications, our explanations target medical professionals who can appropriately interpret physiological measurements, rather than patients without specialized knowledge.

A notable property of IEMM is the generation of inherently shallow decision trees. Following the IMM design principle, IEMM produces exactly one leaf per cluster, limiting tree depth to at most $|\mathbb{F}| - 1$. This structural constraint enhances interpretability—a primary goal of explainable AI—though it may occasionally result in explanations that cannot fully capture complex data patterns, potentially creating overly rigid explainers for certain applications.

Our research inaugurates **perspectives** for future investigation, particularly in two key directions:

- **Elicitation of Utilities in Imprecise Contexts**: While we have proposed a family of natural constructions for utility functions, domain-specific adaptations warrant further exploration in order to better capture the preferences of decision-makers. Future work could focus on developing systematic methods for characterizing and eliciting these utilities.
- **Advanced interpretable evidential classifiers**: Developing more sophisticated interpretable evidential classifiers that exceed the performance of standard decision trees represents a significant opportunity. Potential approaches include incorporating other DT-based explainers into the evidential case Lawless & Gunluk (2022); Fleissner et al. (2024) and constructing Belief Rule-Based Jiao et al. (2015) Explainers, which can incorporate the "non-categoricalness" of the original evidential partition into the explanation. Additionally, extending these methods to better account for the open-world hypothesis could enhance their robustness in real-world applications.

In conclusion, by advancing methods for cautious and explainable clustering, our work contributes to the broader goal of developing AI systems that effectively handle uncertainty while remaining interpretable to human experts. The IEMM algorithm and its theoretical foundations represent a step toward AI systems that acknowledge imperfect information, incorporate domain expertise, and communicate their reasoning in an accessible manner—all key requirements for responsible AI deployment in high-stakes decision-making contexts.

REFERENCES

Stefan Aeberhard, Dionysius Coomans, and Olivier de Vel. Comparison of classifiers in high dimensional settings. In Susan Flockton and Terence Payne (eds.), *Machine Learning: ECML-91*, volume 482 of *Lecture Notes in Computer Science*, pp. 177–186. Springer, 1991. doi: 10.1007/3-540-54781-8_17.

Miguel Alvarez-Garcia, Raquel Ibar-Alonso, and Mar Arenas-Parra. A comprehensive framework for explainable cluster analysis. *Information Sciences*, 663:120282, March 2024. ISSN 00200255. doi: 10.1016/j.ins.2024.120282.

Leila Amgoud and Jonathan Ben-Naim. Axiomatic Foundations of Explainability. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 636–642, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/90.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty Sets for Image Classifiers using Conformal Prediction, September 2022.

Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the explanatory power of Boolean decision trees. *Data & Knowledge Engineering*, 142:102088, November 2022. ISSN 0169023X. doi: 10.1016/j.datak.2022.102088.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.*, 11: 1803–1831, August 2010. ISSN 1532-4435.

Sayan Bandyapadhyay, Fedor V. Fomin, Petr A. Golovach, William Lochet, Nidhi Purohit, and Kirill Simonov. How to find a good explanation for clustering? *Artificial Intelligence*, 322: 103948, September 2023. ISSN 00043702. doi: 10.1016/j.artint.2023.103948.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020. ISSN 15662535. doi: 10.1016/j.inffus.2019.12.012.

Asa Ben-Hur, David Horn, Hava T Siegelmann, and Vladimir Vapnik. Support Vector Clustering. *Journal of Machine Learning Research*, 2001.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation. *Advances in Neural Information Processing Systems*, 35: 29205–29216, December 2022.

Andrea Campagner, Davide Ciucci, and Thierry Denœux. A general framework for evaluating and comparing soft clusterings. *Information Sciences*, 623:70–93, April 2023. ISSN 0020-0255. doi: 10.1016/j.ins.2022.11.114.

Emilio Carrizosa, Kseniia Kurishchenko, Alfredo Marín, and Dolores Romero Morales. Interpreting clusters via prototype optimization. *Omega*, 107:102543, February 2022. ISSN 0305-0483. doi: 10.1016/j.omega.2021.102543.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, August 2019. ISSN 2079-9292. doi: 10.3390/electronics8080832.

T. Denoeux and M.S. Bjanger. Induction of decision trees from partially classified data using belief functions. In *SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. 'Cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions' (Cat. No.00CH37166)*, volume 4, pp. 2923–2928, Nashville, TN, USA, 2000. IEEE. ISBN 978-0-7803-6583-4. doi: 10.1109/ICSMC.2000.884444.

Thierry Denoux, Shoumei Li, and Songsak Sriboonchitta. Evaluating and Comparing Soft Partitions: An Approach Based on Dempster–Shafer Theory. *IEEE Transactions on Fuzzy Systems*, 26(3):1231–1244, June 2018. ISSN 1063-6706, 1941-0034. doi: 10.1109/TFUZZ.2017.2718484.

Didier Dubois and Henri Prade. Formal Representations of Uncertainty. In Denis Bouyssou, Didier Dubois, Marc Pirlot, and Henri Prade (eds.), *Decision-making Process*, pp. 85–156. Wiley, 1 edition, January 2009. ISBN 978-1-84821-116-2 978-0-470-61187-6. doi: 10.1002/9780470611876.ch3.

Didler Dubois and Henri Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4(3):244–264, 1988. ISSN 1467-8640. doi: 10.1111/j.1467-8640.1988.tb00279.x.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.

Charles A. Ellis, Mohammad S. E. Sendi, Eloy P. T. Geenjaar, Sergey M. Plis, Robyn L. Miller, and Vince D. Calhoun. Algorithm-Agnostic Explainability for Unsupervised Clustering, August 2021.

Charles A. Ellis, Robyn L. Miller, and Vince D. Calhoun. Explainable fuzzy clustering framework reveals divergent default mode network connectivity dynamics in schizophrenia. *Frontiers in Psychiatry*, 15:1165424, February 2024. ISSN 1664-0640. doi: 10.3389/fpsyt.2024.1165424.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7: 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Maximilian Fleissner, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Explaining Kernel Clustering via Decision Trees, February 2024.

Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. ExKMC: Expanding Explainable k-Means Clustering, July 2020.

Magzhan Gabidolla and Miguel Á. Carreira-Perpiñán. Optimal Interpretable Clustering Using Oblique Decision Trees. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 400–410, Washington DC USA, August 2022. ACM. ISBN 978-1-4503-9385-0. doi: 10.1145/3534678.3539361.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5): 93:1–93:42, August 2018. ISSN 0360-0300. doi: 10.1145/3236009.

Pierre Hansen and Brigitte Jaumard. Cluster analysis and mathematical programming. *Mathematical Programming*, 79(1):191–215, October 1997. ISSN 1436-4646. doi: 10.1007/BF02614317.

J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, 28(1):100, 1979. ISSN 00359254. doi: 10.2307/2346830.

Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois, and Yolande Le Gall. Evidential Random Forests. *Expert Systems with Applications*, 230:120652, November 2023a. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.120652.

Arthur Hoarau, Constance Thierry, Arnaud Martin, Jean-Christophe Dubois, and Yolande Le Gall. Datasets with Rich Labels for Machine Learning. In *FUZZ*, January 2023b.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3.

Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 548–557. PMLR, August 2022.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-Based Explanations for Machine Learning Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):1511–1519, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33011511.

Abdelhak Imoussaten. The study of the hyper-parameter modelling the decision rule of the cautious classifiers based on the F-beta measure. *Array*, 19:100310, September 2023. ISSN 2590-0056. doi: 10.1016/j.array.2023.100310.

Abdelhak Imoussaten. Convex Mixture Criterion Based Evidential Set-Valued Classification. In *FUZZ IEEE 2025 - IEEE International Conference on Fuzzy Systems*. IEEE, July 2025. doi: 10.1109/FUZZ62266.2025.11152101.

Abdelhak Imoussaten and Lucie Jacquin. Cautious classification based on belief functions theory and imprecise relabelling. *International Journal of Approximate Reasoning*, 142:130–146, March 2022. ISSN 0888-613X. doi: 10.1016/j.ijar.2021.11.009.

Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On Tackling Explanation Redundancy in Decision Trees. *Journal of Artificial Intelligence Research*, 75:261–321, September 2022a. ISSN 1076-9757. doi: 10.1613/jair.1.13575.

Yacine Izza, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. Provably Precise, Succinct and Efficient Explanations for Decision Trees, May 2022b.

Lucie Jacquin, Abdelhak Imoussaten, François Trousset, Jacky Montmain, and Didier Perrin. Evidential Classification of Incomplete Data via Imprecise Relabelling: Application to Plastic Sorting. In Nahla Ben Amor, Benjamin Quost, and Martin Theobald (eds.), *Scalable Uncertainty Management*, volume 11940, pp. 122–135. Springer International Publishing, Cham, 2019. ISBN 978-3-030-35513-5 978-3-030-35514-2. doi: 10.1007/978-3-030-35514-2_10.

Lianmeng Jiao, Quan Pan, Thierry Denœux, Yan Liang, and Xiaoxue Feng. Belief rule-based classification system: Extension of FRBCS in belief functions framework. *Information Sciences*, 309: 26–49, July 2015. ISSN 00200255. doi: 10.1016/j.ins.2015.03.005.

Anne-Laure Jousselme and Patrick Maupin. Distances in evidence theory: Comprehensive survey and generalizations. *International Journal of Approximate Reasoning*, 53(2):118–145, February 2012. ISSN 0888-613X. doi: 10.1016/j.ijar.2011.07.006.

Anne-Laure Jousselme, Dominic Grenier, and Éloi Bossé. A new distance between two bodies of evidence. *Information Fusion*, 2(2):91–101, June 2001. ISSN 15662535. doi: 10.1016/S1566-2535(01)00026-4.

Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, July 1993. ISBN 978-0-521-43883-4.

R. Krishnapuram and J.M. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110, May 1993. ISSN 1941-0034. doi: 10.1109/91.227387.

Lucie Kunitomo-Jacquin, Abdelhak Imoussaten, Sebastien Destercke, Christophe Marsala, and Ken Fukuda. Towards evaluating set-valued predictions with partial observations. In *2025 IEEE International Conference on Fuzzy Systems (FUZZ)*, pp. 1–4, July 2025. doi: 10.1109/FUZZ62266.2025.11152051.

Eduardo Laber, Lucas Murtinho, and Felipe Oliveira. Shallow decision trees for explainable k-means clustering, August 2022.

Connor Lawless and Oktay Gunluk. Cluster Explanation via Polyhedral Descriptions, October 2022.

Pawan Lingras and Chad West. Interval Set Clustering of Web Users with Rough K-Means. *Journal of Intelligent Information Systems*, 23(1):5–16, July 2004. ISSN 1573-7675. doi: 10.1023/B:JIIS.0000029668.88665.1a.

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles, March 2019.

Liyao Ma and Thierry Denœux. Partial classification in the belief function framework. *Knowledge-Based Systems*, 214:106742, February 2021. ISSN 09507051. doi: 10.1016/j.knosys.2021.106742.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5.1, pp. 281–298. University of California Press, January 1967.

Konstantin Makarychev and Liren Shan. Explainable k-means. Don't be greedy, plant bigger trees!, April 2022.

Marie-Hélène Masson and T. Denœux. ECM: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397, April 2008. ISSN 00313203. doi: 10.1016/j.patcog.2007.08.014.

Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-Means and k-Medians Clustering. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7055–7065. PMLR, November 2020.

Nina Narodytska, Aditya Shrotri, Kuldeep S. Meel, Alexey Ignatiev, and Joao Marques-Silva. Assessing Heuristic Machine Learning Explanations with Model Counting. In Mikoláš Janota and Inês Lynce (eds.), *Theory and Applications of Satisfiability Testing – SAT 2019*, pp. 267–278, Cham, 2019. Springer International Publishing. ISBN 978-3-030-24258-9. doi: 10.1007/978-3-030-24258-9_19.

Vu-Linh Nguyen, Sébastien Destercke, Marie-Hélène Masson, and Eyke Hüllermeier. Reliable Multi-class Classification based on Pairwise Epistemic and Aleatoric Uncertainty. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 5089–5095, 2018.

Zdzisław Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11(5):341–356, October 1982. ISSN 1573-7640. doi: 10.1007/BF01001956.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, September 1987. ISSN 0020-7373. doi: 10.1016/S0020-7373(87)80053-6.

William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. ISSN 0162-1459. doi: 10.2307/2284239.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

Lior Rokach and Oded Maimon. Decision Trees. In Oded Maimon and Lior Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, pp. 165–192. Springer-Verlag, New York, 2005. ISBN 978-0-387-24435-8. doi: 10.1007/0-387-25465-X_9.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x.

Enrique H. Ruspini. A new approach to clustering. *Information and Control*, 15(1):22–32, July 1969. ISSN 0019-9958. doi: 10.1016/S0019-9958(69)90591-9.

Leonard J. Savage. *The Foundations of Statistics*. John Wiley and Sons, New York, 1954. ISBN 978-0-486-62349-8 978-0-486-13710-0.

David Schmeidler. Subjective Probability and Expected Utility without Additivity. *Econometrica*, 57(3):571–587, 1989. ISSN 0012-9682. doi: 10.2307/1911053.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, April 1976. ISBN 978-0-691-10042-5.

Philippe Smets (ed.). *Non-Standard Logics for Automated Reasoning*. Academic Press, San Diego, 1988.

Philippe Smets. Constructing the Pignistic Probability Function in a Context of Uncertainty. In *Machine Intelligence and Pattern Recognition*, volume 10, pp. 29–39. Elsevier, 1990. ISBN 978-0-444-88738-2. doi: 10.1016/B978-0-444-88738-2.50010-5.

Armel Soubeiga and Violaine Antoine. Evclust: Python library for evidential clustering, February 2025.

Guolong Su, Dennis Wei, Kush R. Varshney, and Dmitry M. Malioutov. Interpretable Two-level Boolean Rule Learning for Classification, November 2015.

Sariel Tutay and Amit Somech. Cluster-Explorer: An interactive Framework for Explaining Black-Box Clustering Results. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5106–5110, Birmingham United Kingdom, October 2023. ACM. ISBN 979-8-4007-0124-5. doi: 10.1145/3583780.3614734.

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior, 2nd Rev. Ed*. Theory of Games and Economic Behavior, 2nd Rev. Ed. Princeton University Press, Princeton, NJ, US, 1947.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Rui Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, May 2005. ISSN 1941-0093. doi: 10.1109/TNN.2005.845141.

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, June 1965. ISSN 0019-9958. doi: 10.1016/S0019-9958(65)90241-X.

Marco Zaffalon, Giorgio Corani, and Denis Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, November 2012. ISSN 0888-613X. doi: 10.1016/j.ijar.2012.06.022.

Haifei Zhang. *Explainable Cautious Classifiers*. PhD thesis, Université de Technologie de Compiègne, November 2023.

Zuowei Zhang, Yiru Zhang, Hongpeng Tian, Arnaud Martin, Zhunga Liu, and Weiping Ding. A survey of evidential clustering: Definitions, methods, and applications. *Information Fusion*, pp. 102736, October 2024. ISSN 15662535. doi: 10.1016/j.inffus.2024.102736.

A Evidential Theory and Clustering

A.1 Evidential Theory

From a mass function, one can derive two key set functions—belief and plausibility:

$$\mathrm{Bel}_m(A) = \sum_{B \subseteq A} m(B) \quad \text{and} \quad \mathrm{Pl}_m(A) = \sum_{B \cap A \neq \varnothing} m(B).$$

The belief function $\mathrm{Bel}_m(A)$ for $A \subseteq \Omega$ represents the degree of confidence that 'the correct hypothesis ω belongs to A'. In contrast, the plausibility function $\mathrm{Pl}_m(A)$ captures the degree of confidence that 'it is not impossible for the correct hypothesis ω to belong to A'. Mass functions generalize probability mass functions by distributing belief across all subsets of Ω, rather than only its individual elements.

Another useful measure is the **pignistic probability** Smets (1990), which transforms a mass function into a probability distribution over Ω:

$$\mathrm{BetP}_m(\omega) = \sum_{\{\omega\} \subseteq A \subseteq \Omega} \frac{m(A)}{|A|}.$$

The pignistic probability BetP_m can be interpreted as providing the best Bayesian approximation to the mass function m.

Similarly, there are several techniques to derive the categorical mass closest to a given evidential clustering function Imoussaten (2025). A prominent method relies on the **strong dominance** criterion: for any $\omega, \omega' \in \Omega$, ω strongly dominates ω' if and only if $\mathrm{Bel}_m(\{\omega\}) > \mathrm{Pl}_m(\{\omega'\})$. Mapping m to the set of non-strongly dominated clusters yields a categorical mass function.

A.2 Clustering Functions

Figure 5 illustrates various clustering methods for $\Omega = \{\omega_1, \omega_2\}$. While hard clustering assigns each point to exactly one cluster, evidential clustering offers a more sophisticated representation by capturing uncertainty and imprecision. It identifies points that may plausibly belong to multiple clusters (represented in the figure by those primarily associated with $\omega_1 \cup \omega_2$) and accounts for points not clearly associated with any cluster (shown as those predominantly linked to $\varnothing$).

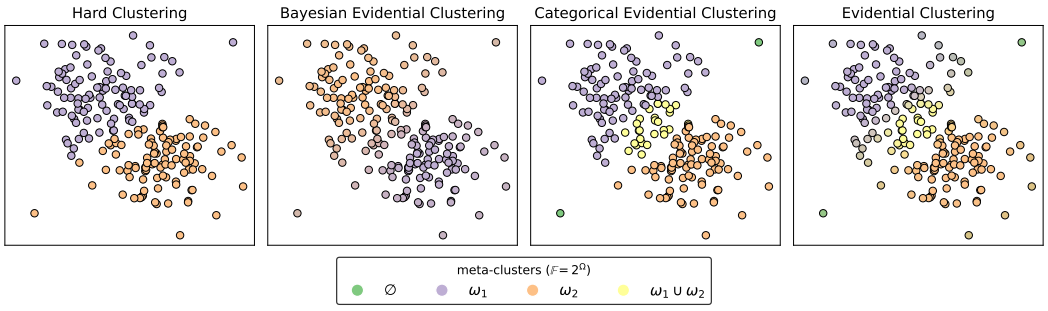

Figure 5: A representation of different clustering functions over a synthetic dataset. In this case, $\mathbb{X} = \mathbb{R}^2$ and $|\Omega| = 2$. Dataset was constructed by sampling 100 points from two normal distributions with centers at $(3, 5)$ and $(5, 3)$ and σ of 1. Two outliers were added, at $(2, 2)$ and $(6, 6)$. The `evclust` package Soubeiga & Antoine (2025) was used to perform clustering. Hard clustering assigns each point to a single cluster. Bayesian evidential clustering gives a membership level for each observation. Categorical evidential clustering introduces the information about in-between points ($\omega_1 \cup \omega_2$) and outliers ($\varnothing$). Finally, evidential clustering combines all the previous information. The gradient of colors for each point visually represents the mass.

B On Simplification Explanation Techniques

Simplification techniques typically rely on rule extraction methods, encompassing both global and local approaches. Studies have been conducted to assess the quality of these explanations Amgoud &

Ben-Naim (2022). Below, we introduce definitions that characterize effective explanations produced by simplification techniques.

A **feature literal** is a pair $\langle \mathcal{A}, v \rangle$ where $\mathcal{A} \in \mathcal{D}$ and $v \in \mathcal{A}$. Let $\mathbf{L}$ be the set of all feature literals for all attributtes. A consistent subset of feature literals is $L \subset \mathbf{L}$ such that $\langle \mathcal{A}, v \rangle, \langle \mathcal{A}, v' \rangle \in L \Rightarrow v = v'$. Let $\mathbf{C} \subseteq 2^{\mathbf{L}}$ be the set of **all consistent subsets of feature literals**. Each $\mathbf{D} \subset \mathbf{C}$ induces a map $\mathrm{DNF} : \mathbb{X} \to \{\mathrm{True}, \mathrm{False}\}$ with

$$\mathrm{DNF}_{\mathbf{D}}(x) = \bigvee_{L \in \mathbf{D}} \left(\bigwedge_{\langle \mathcal{A}, v \rangle \in L} (x_{\mathcal{A}} = v) \right) \tag{6}$$

which is a Disjunctive Normal Form (DNF) Su et al. (2015). A DNF can serve as a human-interpretable classification model. When a DNF matches the behavior of a black-box classifier, we achieve a particularly desirable outcome known as an abductive explanation.

A Concrete Example: Consider a philosopher studying living beings who observes two key characteristics: their appearance and their mode of locomotion. To formalize this classification problem, the philosopher defines the feature space of conceivable living beings as $\mathbb{X} = \texttt{App} \times \texttt{Move}$, where

$$\texttt{App} = \{\texttt{feathered}, \texttt{featherless}\} \text{ and } \texttt{Move} = \{\texttt{biped}, \texttt{non-biped}\}.$$

From these features, we can construct the set of feature literals:

$$\mathbf{L} = \{\langle \texttt{App}, \texttt{feathered} \rangle, \langle \texttt{App}, \texttt{featherless} \rangle, \langle \texttt{Move}, \texttt{biped} \rangle, \langle \texttt{Move}, \texttt{non-biped} \rangle\}.$$

The set of all consistent subsets of feature literals encompasses all possible combinations that do not contain contradictory values for the same attribute:

$$\begin{aligned}\mathbf{C} = \{&\varnothing,\\ &\{\langle \texttt{App}, \texttt{feathered} \rangle\}, \{\langle \texttt{App}, \texttt{featherless} \rangle\}, \{\langle \texttt{Move}, \texttt{biped} \rangle\}, \{\langle \texttt{Move}, \texttt{non-biped} \rangle\},\\ &\{\langle \texttt{App}, \texttt{feathered} \rangle, \langle \texttt{Move}, \texttt{biped} \rangle\}, \{\langle \texttt{App}, \texttt{feathered} \rangle, \langle \texttt{Move}, \texttt{non-biped} \rangle\},\\ &\{\langle \texttt{App}, \texttt{featherless} \rangle, \langle \texttt{Move}, \texttt{biped} \rangle\}, \{\langle \texttt{App}, \texttt{featherless} \rangle, \langle \texttt{Move}, \texttt{non-biped} \rangle\}\\ &\}.\end{aligned}$$

An example of a DNF is given by $\mathbf{D} = \{\{\langle \texttt{App}, \texttt{featherless} \rangle, \langle \texttt{Move}, \texttt{biped} \rangle\}\}$, which corresponds to featherless bipedal beings. That is, for any living being x, we have $\mathrm{DNF}_{\mathbf{D}}(x) = (x_{\texttt{App}} = \texttt{featherless}) \wedge (x_{\texttt{Move}} = \texttt{biped})$. $\mathrm{DNF}_{\mathbf{D}}(x)$ is true whenever x is a human being.
Conversely, $\mathbf{D}' = \{\{\langle \texttt{App}, \texttt{feathered} \rangle\}, \{\langle \texttt{Move}, \texttt{non-biped} \rangle\}\}$ corresponds to beings that are either feathered or non-bipedal. In this case, the induced DNF is $\mathrm{DNF}_{\mathbf{D}'}(x) = (x_{\texttt{App}} = \texttt{feathered}) \vee (x_{\texttt{Move}} = \texttt{non-biped})$ and $\mathrm{DNF}_{\mathbf{D}'}(x)$ is false whenever x is a human being.

Definition 6. An **abductive explanation** of the label $\omega \in \Omega$ is a $L \in \mathbf{C}$ such that, $\forall x \in \mathbb{X}$,

$$\left(\bigwedge_{\langle \mathcal{A}, v \rangle \in L} (x_{\mathcal{A}} = v) \right) \Rightarrow \Gamma(x) = \omega$$

Abductive explanations were introduced to address the question: "Why is $\Gamma(x) = \omega$?", providing a sufficient reason for characterizing the label ω, where Γ is a supervised classifier Ignatiev et al. (2019). In the context of explainability, an ideal construction would be a system that can provide satisfactory explanations[4] for a classifier's outputs.

Definition 7. An **explainer** of a classifier $\Gamma : \mathbb{X} \to \Omega$ is a map $\chi_\Gamma : \Omega \to 2^{\mathbf{C}}$.

That is, to each class ω, a classifier associates a DNF. If the DNF issued from χ_Γ matches Γ, the classifier provides abductive explanations.

[4]In this work, we consider satisfactory explanations to be "abductive" or "as abductive as possible." However, this might not always be the case. As highlighted in multiple works Barredo Arrieta et al. (2020), the best type of explanation depends on the audience for which this explanation is intended. We develop this discussion further in the conclusion.

C Decision Trees as Explainers

In this section, we provide a brief overview of decision trees (DTs) and their role as explainers. We also establish the relationship between representativity and abductivity in the context of DTs. We adapt the following definition of univariate decision trees from Izza et al. (2022a).

Definition 8. The **graph of a decision tree** $\mathcal{T} = (V, E)$ is a directed acyclic graph in which there is at most one path between any two vertices. The vertex set V is divided into non-terminal vertices N and terminal vertices T, such that $V = N \cup T$. Additionally, $\mathcal{T}$ has a unique root vertex, $\mathrm{root}(\mathcal{T}) \in V$, which has no incoming edges, while every other vertex has exactly one incoming edge.

To each graph of a DT, there are two important associated functions:

- A **split** is a map $\phi : N \to \mathcal{D}$ that assigns an attribute to each non-terminal vertex.
- Let $\mathrm{children}(r) = \{s \in V \mid (r, s) \in E\}$ be the set of children of a vertex r. A **decision** is a map $\varepsilon : E \to \mathbf{L}$ such that, for every non-terminal vertex $r \in N$, there exists a bijection $\varepsilon_r : \mathrm{children}(r) \to \phi(r)$ satisfying $\varepsilon(r, s) = \langle \phi(r), \varepsilon_r(s) \rangle$.

It is well known that any binary decision tree can be transformed in linear time into an equivalent disjunctive normal form (DNF) expression Audemard et al. (2022). This property is often referenced when DTs are described as "interpretable" Guidotti et al. (2018). With this in mind, we associate each vertex with a path, which serves as the foundation for interpreting a decision tree as an explainer.

For a fixed graph of a DT $\mathcal{T}$, let $\mathrm{DNF}(r)$ be the set of literals associated with the edges that link the root to vertex r. All literals in $\mathrm{DNF}(r)$ are consistent Izza et al. (2022a). That is, $\mathrm{DNF} : V \to 2^{\mathbf{C}}$. Let $\mathbf{D} = \mathrm{DNF}(T)$ be the set of all DNFs associated with terminal vertices.

Definition 9. A **path** is a map $\Upsilon : \mathbb{X} \to \mathbf{D}$ such that, for all $x \in \mathbb{X}$,

$$\bigwedge_{\langle \mathcal{A}, v \rangle \in \Upsilon(x)} (x_{\mathcal{A}} = v).$$

The partitioning nature of decision trees ensures that each path is well-defined, meaning every possible observation follows a unique path. This characteristic allows us to interpret vertices as subsets of the feature space Hoarau et al. (2023a).

Definition 10. A **node** is a nonempty subset $S \subseteq \mathbb{X}$.

Every achievable vertex can be trivially associated with a unique node by its DNF. We call **leaves** the nodes associated with terminal vertices. The set $\mathbf{D}$ can be understood as the explanation for each leaf. Associating leaves with explanations allows us to define the DT as a classifier.

Definition 11. A **decision tree** is a map $\Delta : \mathbb{X} \to \Omega$ to which a path Υ_Δ provides an abductive explanation. That is, $\forall x \in \mathbb{X}$,

$$\bigwedge_{\langle \mathcal{A}, v \rangle \in \Upsilon(x)} (x_{\mathcal{A}} = v) \Rightarrow \Delta(x) = \omega.$$

Let, $\forall \omega \in \Omega$, $\mathcal{L}^\Delta_\omega = \{\Upsilon^{-1}(\{L\}) : L \in \Upsilon(\Delta^{-1}(\{\omega\}))\}$ be the set of all leaves associated with the explanation of ω. The **DT explainer** χ^Δ_Γ associated with Δ is an explainer that, for any label, returns all paths explaining it. That is, $\chi^\Delta_\Gamma(\omega) = \Upsilon_\Delta[\mathcal{L}^\Delta_\omega] = \{\Upsilon_\Delta(x) : x \in \mathcal{L}^\Delta_\omega\}$.

Our investigation focuses on the quality of explanations when, in the context of model simplification, the original classifier diverges from the decision tree explaining it. We borrow the concept of representative explainer from Amgoud & Ben-Naim (2022). A representative explainer is one that, for all observations x with label $\omega = \Gamma(x)$, can provide an explanation L that holds at x. That is, there exists a set of literals $L \in \chi_\Gamma(\omega)$ such that $\langle \mathcal{A}, v \rangle \in L \Rightarrow x_{\mathcal{A}} = v$.

Definition 12. A **representative explainer** is an explainer χ_Γ such that, $\forall \omega \in \Omega$, $\forall x \in \Gamma^{-1}(\{\omega\})$, $\exists L \in \chi_\Gamma(\omega)$ such that, $\forall \langle \mathcal{A}, v \rangle \in L, x_{\mathcal{A}} = v$.

The work in Amgoud & Ben-Naim (2022) proves that every explainer providing abductive explanations is representative. We complement this result by proving that every representative DT explainer provides abductive explanations. Thus, for DT explainers, representativity and abductivity are equivalent.

Theorem 1. *If the χ^{Δ}_{Γ} explainer is representative, it provides abductive explanations.*

Proof. We proceed by contradiction. Assume the DT explainer does not provide abductive explanations.

From definition 11, this implies that $\Gamma \neq \Delta$. That is, there exists $x \in \mathbb{X}$ such that $\omega_\Gamma = \Gamma(x) \neq \Delta(x) = \omega_\Delta$. Since the leaves form a partition of the feature space and $x \in \mathcal{L}^{\Delta}_{\omega_\Delta}$, we have $\Upsilon_\Delta(x) \notin \Upsilon_\Delta[\mathcal{L}^{\Delta}_{\omega_\Gamma}]$, and the explainer is not representative. □

D On Explanation Costs

In the context of a categorical evidential partition $\overline{\mathcal{M}}_c$, we want to characterize the cost of explaining a point x with a cautious explainer induced by some interpretable classifier $\Delta : X \to 2^{\Omega}$.

The utility $\mathcal{U}(A, \overline{\mathcal{M}}_c(x))$ quantifies the satisfaction of assigning metacluster A to observation x and, therefore, equals the cost of not assigning A to x.

Thus, the total cost can be understood as the function $\overline{\text{Cost}_{\mathcal{M}_c,\Delta}} : X \to [0, |\mathbb{F}| - 1]$, that maps x to the sum of costs from not assigning x to all metaclusters $A \neq \Delta(x)$,

$$\overline{\text{Cost}_{\mathcal{M}_c,\Delta}}(x) = \sum_{A \neq \Delta(x)} \mathcal{U}(A, \overline{\mathcal{M}}_c(x)). \tag{7}$$

Conversely, $1 - \mathcal{U}(A, \overline{\mathcal{M}}_c(x))$ represents the cost of assigning A to x, leading to $\underline{\text{Cost}_{\mathcal{M}_c,\Delta}} : X \to [0, 1]$, an alternative expression for total cost:

$$\underline{\text{Cost}_{\mathcal{M}_c,\Delta}}(x) = 1 - \mathcal{U}(\Delta(x), \overline{\mathcal{M}}_c(x)). \tag{8}$$

It always holds that $\underline{\text{Cost}_{\mathcal{M}_c,\Delta}}(x) \leq \overline{\text{Cost}_{\mathcal{M}_c,\Delta}}(x)$. When $\mathcal{U}(A, B) = \mathbb{1}_{A=B}$, the equality holds. Furthermore, if $\mathcal{M}_c$ induces a hard clustering and Δ is IMM-like (with exactly one centroid per leaf), both equal one (and not zero) if and only if x is a mistake as stated in Definition 2.

D.1 A simple example

Figure 6 and Table 1 illustrate the utility concept through a concrete example. The value $\mathcal{U}(\{\omega_1, \omega_2\}, \{\omega_1\})$ quantifies how tolerable the mistake at x_1 is. When $\mathcal{U}(\{\omega_1, \omega_2\}, \{\omega_1\}) = 0$, this mistake becomes as intolerable as the one at x_2. Conversely, when $\mathcal{U}(\{\omega_1, \omega_2\}, \{\omega_1\}) = 1$, it becomes as tolerable as the correct assignment at x_0. In this latter scenario, removing x_2 from the dataset would yield an optimal assignment.

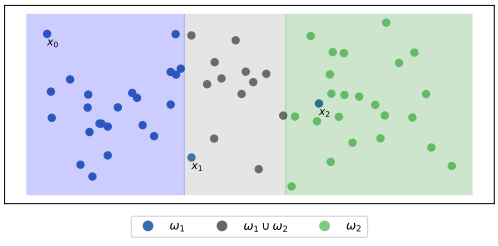

Figure 6: Illustration of a categorical evidential classifier and space partition in $\mathbb{X} = \mathbb{R}^2$. The partition Δ separates x_1 and x_2 from their respective metaclusters, while correctly assigning all other observations.

x	x_0	x_1	x_2
$\Delta(x)$	$\{\omega_1\}$	$\{\omega_1, \omega_2\}$	$\{\omega_2\}$
$\underline{\text{Cost}_{\mathcal{M}_c,\Delta}}(x)$	0	$1 - \mathcal{U}(\{\omega_1, \omega_2\}, \{\omega_1\})$	1
$2^{\Omega} \setminus \Delta(x)$	$\{\{\omega_2\}, \{\omega_1, \omega_2\}\}$	$\{\{\omega_1\}, \{\omega_2\}\}$	$\{\{\omega_1\}, \{\omega_1, \omega_2\}\}$
$\overline{\text{Cost}_{\mathcal{M}_c,\Delta}}(x)$	$\mathcal{U}(\{\omega_1, \omega_2\}, \{\omega_1\})$	1	$1 + \mathcal{U}(\{\omega_1, \omega_2\}, \{\omega_1\})$

Table 1: For each point highlighted in Figure 6, we present the point x, the metacluster assigned by classifier Δ, the cost of assigning x to the metacluster designated by Δ, the set of metaclusters not assigned by Δ, and the cost of not assigning x to them. Note that $\overline{\mathcal{M}}_c(x_0) = \overline{\mathcal{M}}_c(x_1) = \overline{\mathcal{M}}_c(x_2) = \{\omega_1\}$.

E Relating Mistakeness and Representativeness

In this section, we show that the evidential representativeness and the total evidential mistakeness (the sum of the evidential mistakeness of each leaf) are equivalent in terms of measuring the quality of a IMM-like decision tree.

Theorem 2. *Let $\Delta, \Delta' : \mathbb{X} \to 2^\Omega$ be two IMM-like decision trees. Then, for any evidential partition $\mathcal{M}$ and utility $\mathcal{U}$,*

$$\begin{aligned} & \mathcal{R}_{\mathcal{M},\mathcal{U}}(\Delta) \geq \mathcal{R}_{\mathcal{M},\mathcal{U}}(\Delta') \\ \iff & \sum_{A\subset\Omega} \overline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^\Delta) \leq \sum_{A\subset\Omega} \overline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^{\Delta'}) \\ \iff & \sum_{A\subset\Omega} \underline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^\Delta) \leq \sum_{A\subset\Omega} \underline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^{\Delta'}) \end{aligned}$$

where $v_A \in \mathcal{L}_A^\Delta$ which is the leaf associated with the cluster A in the decision tree Δ.

Proof. We start by establishing the relation between the two mistakenness functions. By definition, $x \in \mathcal{L}_A^\Delta \iff \Delta(x) = A$. From equations (4) and (5),

$$\begin{aligned} \overline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^\Delta) &= \sum_{x\in\mathcal{L}_A^\Delta} \sum_{\Delta(x)\neq C} \sum_{B\in\mathbb{F}_\mathcal{M}} \mathcal{U}(C,B) m_x(B), \\ \underline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^\Delta) &= \sum_{x\in\mathcal{L}_A^\Delta} \sum_{B\in\mathbb{F}_\mathcal{M}} (1-\mathcal{U}(A,B)) m_x(B). \end{aligned}$$

Then,

$$\begin{aligned} & \sum_{A\subset\Omega} \underline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^\Delta) - \sum_{A\subset\Omega} \overline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^\Delta) \\ &= \sum_{A\subset\Omega} \sum_{x\in\mathcal{L}_A^\Delta} \sum_{B\in\mathbb{F}_\mathcal{M}} m_x(B) \left((1-\mathcal{U}(A,B)) - \sum_{A\neq C} \mathcal{U}(C,B) \right) \\ &= \sum_{x\in X} \sum_{B\in\mathbb{F}_\mathcal{M}} m_x(B) \left(1 - \sum_{C\subset\Omega} \mathcal{U}(C,B) \right) \\ &= |X| - \sum_{x\in X} \sum_{B\in\mathbb{F}_\mathcal{M}} m_x(B) \left(\sum_{C\subset\Omega} \mathcal{U}(C,B) \right) = |X| - \kappa_{\mathcal{M},\mathcal{U}}. \end{aligned}$$

where $\kappa_{\mathcal{M},\mathcal{U}}$ is a constant that depends only on the evidential partition $\mathcal{M}$ and utility $\mathcal{U}$, but not on the specific decision tree Δ.

Also, from equation (3),

$$|X|\mathcal{R}_{\mathcal{M},\mathcal{U}}(\Delta) = \sum_{A\subset\Omega} \sum_{x\in\mathcal{L}_A^\Delta} \sum_{B\in\mathbb{F}_\mathcal{M}} \mathcal{U}(\Delta(x),B) m_x(B).$$

Similarly,

$$\begin{aligned} & |X|\mathcal{R}_{\mathcal{M},\mathcal{U}}(\Delta) + \sum_{A\subset\Omega} \overline{M}_{\mathcal{M},\mathcal{U}}(\mathcal{L}_A^\Delta) = \\ & \sum_{A\subset\Omega} \sum_{x\in\mathcal{L}_A^\Delta} \sum_{B\in\mathbb{F}_\mathcal{M}} m_x(B) \left(\mathcal{U}(A,B) + \sum_{A\neq C} \mathcal{U}(C,B) \right) = \kappa_{\mathcal{M},\mathcal{U}}. \end{aligned}$$

Since all three measures are related by affine transformations with the same constant terms, they preserve the same ordering relationships between different decision trees. Therefore, comparing two trees Δ and Δ' using any of these measures yields equivalent results, proving the stated equivalences. □

F CHOOSING A UTILITY FUNCTION

When an explainer yields a metacluster A, while the original classifier assigns B, two types of errors can occur. The first is insufficient coverage, measured by $|A^C \cap B|$ - where the explainer fails to include all elements of the true metacluster. The second is excessive coverage, measured by $|A \cap B^C|$ - where the explainer includes elements not in the true metacluster. Penalizing insufficient coverage indicates the explainer is not cautious enough, while penalizing excessive coverage suggests it is too cautious.

To address both error types, we introduce two families of utility functions with a positive parameter λ controlling their behavior:

$$\overline{\mathcal{U}}^{\lambda}(A,B) = \left(\frac{|A \cap B|}{|A \cup B|}\mathbb{1}_{B \subset A}\right)^{1/\lambda} \text{ and } \underline{\mathcal{U}}^{\lambda}(A,B) = \left(\frac{|A \cap B|}{|A \cup B|}\mathbb{1}_{A \subset B}\right)^{1/\lambda}.$$

These utility functions exhibit complementary tolerance behaviors. The function $\overline{\mathcal{U}}^{\lambda}(A,B)$ assigns zero utility when $A^C \cap B \neq \varnothing$, making it completely intolerant to insufficient coverage while allowing parameter λ to modulate tolerance to excessive coverage. Higher λ values reduce penalties for excessive coverage, embodying a more cautious approach. In contrast, $\underline{\mathcal{U}}^{\lambda}(A,B)$ assigns zero utility when $A \cap B^C \neq \varnothing$, showing complete intolerance to excessive coverage while λ controls the degree of tolerance to insufficient coverage—higher λ values reducing penalties for insufficient coverage and representing a more risk-attractive approach.

These combine into a comprehensive family of utility functions for $\lambda \in \mathbb{R}^*$:

$$\mathcal{U}^{\lambda}(A,B) = \begin{cases} \underline{\mathcal{U}}^{|\lambda|}(A,B) & \text{if } \lambda < 0 \\ \overline{\mathcal{U}}^{|\lambda|}(A,B) & \text{if } \lambda > 0 \end{cases}.$$

We also define special cases as limits when λ approaches 0 and $\pm\infty$. That gives $\mathcal{U}^{0}(A,B) = \mathbb{1}_{A=B}$, $\mathcal{U}^{-\infty}(A,B) = \mathbb{1}_{A \subset B}$ and $\mathcal{U}^{\infty}(A,B) = \mathbb{1}_{B \subset A}$.

Finally, based on these, we define the **λ-evidential mistakeness** as:

$$M_{\mathcal{M}}^{\lambda} = \begin{cases} \overline{M}_{\mathcal{M},\mathcal{U}^{\lambda}} & \text{if } \lambda \geq 0 \\ \underline{M}_{\mathcal{M},\mathcal{U}^{\lambda}} & \text{if } \lambda < 0 \end{cases} \tag{9}$$

for any $\lambda \in \mathbb{R} \cup \{\pm\infty\}$. The higher the λ, the more the mistakeness function represents a conservative approach. When the underlying clustering function is hard and $\lambda = 0$ is chosen, the algorithm 1 operates identically to IMM because evidential mistakeness equals the number of mistakes. For cautious partitions as input, varying λ controls the "level of cautiousness" of the resulting explainer. Figure 7 illustrates how different λ values influence the partitioning of the feature space by IEMM.

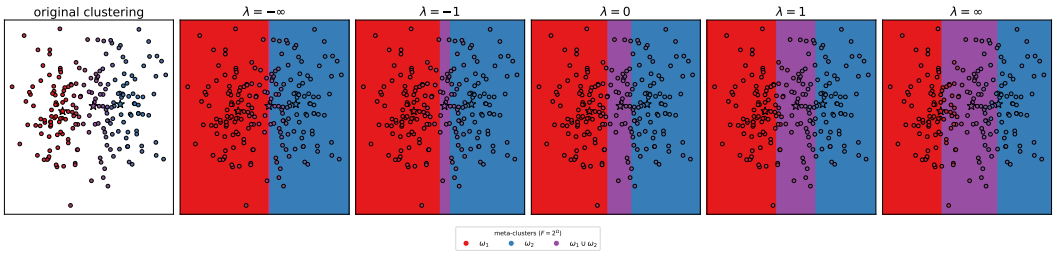

Figure 7: An example, based on a given two-features clustering (column 1), of IEMM partitioning the feature space. The stars represent the centroid of each metacluster. The utility function strongly influences the resulting explanations. The higher the λ, the more the explainer assigns larger portions of the space to metaclusters representing doubt. At the limit, $\lambda = -\infty$ (column 2), the obtained explanations give the maximum possible space to the singleton metaclusters. Conversely, $\lambda = \infty$ (column 6) assigns the maximum possible space to the metaclusters representing doubt.

G EXPERIMENTS

This section presents additional experimental results that validate the IEMM algorithm. We have implemented IEMM using Python 3.13.6. All code is available at OMMITED TO AVOID IDENTIFICATION. The implementation of a decision tree accepting evidential labels was based on the code made available by Hoarau et al. (2023a).

G.1 TESTS ON SYNTHETIC DATASETS

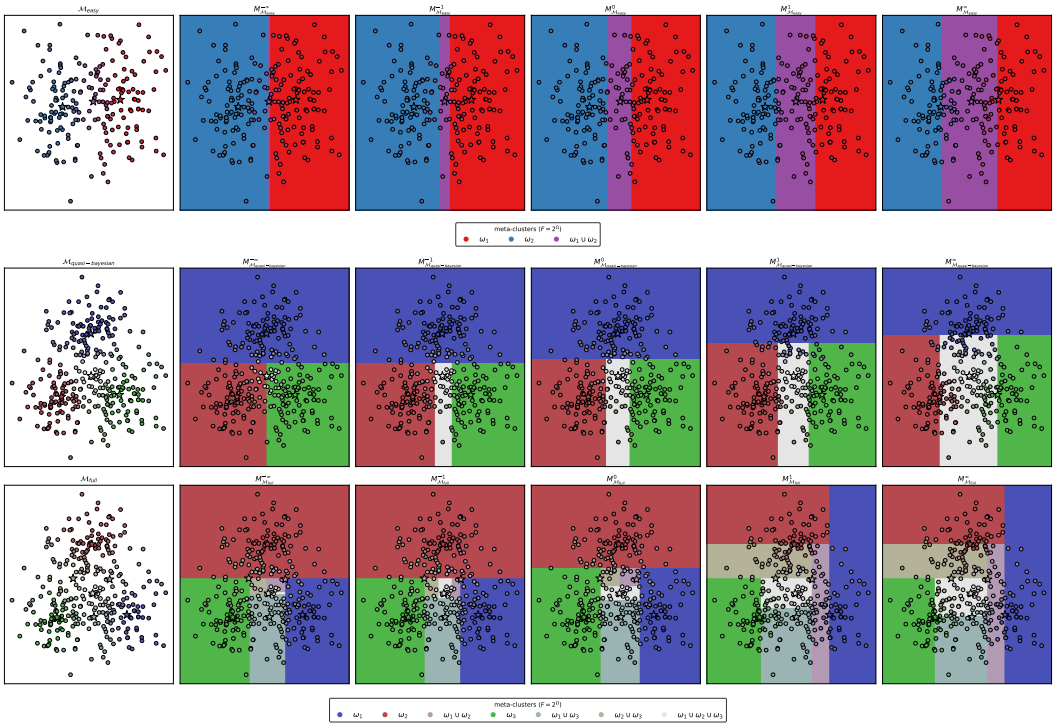

Figure 8: The results of the IEMM on the synthetic dataset for the evidential clustering functions $\mathcal{M}_{\text{easy}}$, $\mathcal{M}_{\text{full}}$, $\mathcal{M}_{\text{quasi-bayesian}}$ and different utility functions. The star represents the centroid of each metacluster. The utility function strongly influences the resulting explanations. The higher the λ, the more the λ-evidential mistakeness function assigns larger portions of the space to metaclusters representing doubt. At the limit, $\lambda = -\infty$ (column 2), the obtained explanations give the maximum possible space to the singleton metaclusters. Conversely, $\lambda = \infty$ (column 6) assigns the maximum possible space to the metaclusters representing doubt.

Using the `evclust` library Soubeiga & Antoine (2025), we generated three evidential partitions over synthetic datasets of 2 features (x and y). Those were:

- A dataset of 200 entries over which we defined $\mathcal{M}_{\text{easy}}$, with $\Omega = \{\omega_1, \omega_2\}$ and $\mathbb{F}_{\mathcal{M}_{\text{easy}}} = 2^{\Omega} \setminus \varnothing$.
- A dataset of 300 samples and, for $\Omega = \{\omega_1, \omega_2, \omega_3\}$, we generated two types of evidential clustering functions:
  - $\mathcal{M}_{\text{full}}$, an evidential clustering with $\mathbb{F}_{\mathcal{M}_{\text{full}}} = 2^{\Omega} \setminus \varnothing$.
  - $\mathcal{M}_{\text{quasi-bayesian}}$, an evidential clustering that is a quasi-bayesian clustering function. This means that the focal sets are the singletons and the whole space. That is, $\mathbb{F}_{\mathcal{M}_{\text{quasi-bayesian}}} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_1, \omega_2, \omega_3\}\}$.

Then, for each evidential clustering function, we constructed a decision tree using IEMM and the λ-evidential mistakeness for different values of λ. A compact overview of results across utilities and λ is provided in Figure 9, while the conflict/non-specificity trends are summarized in Figure 10. The partition of the space induced by these explanations is illustrated in Figure 8. The decision

tree explainer for $\mathcal{M}_{\text{full}}$ and $\lambda = 0$ is shown in Figure 11. The resulting explanations for the quasi-bayesian clustering function are in Table 2.

Table 3 presents the representativeness achieved in each scenario. Notably, the decision tree generated by fixing $\lambda = \infty$ over the $\mathcal{M}_{\text{easy}}$ dataset achieves the highest representativeness observed, surpassing 93%. This can be interpreted as the expected explanation accuracy.

Across both synthetic and real-world datasets, decision trees obtained using the λ-evidential mistakeness function consistently achieve the best performance in terms of $\mathcal{U}^{\lambda}$-evidential representativeness. The gap between the $\mathcal{U}^{\lambda}$-evidential representativeness and 1 quantifies the loss of accuracy or cost of the explanation.

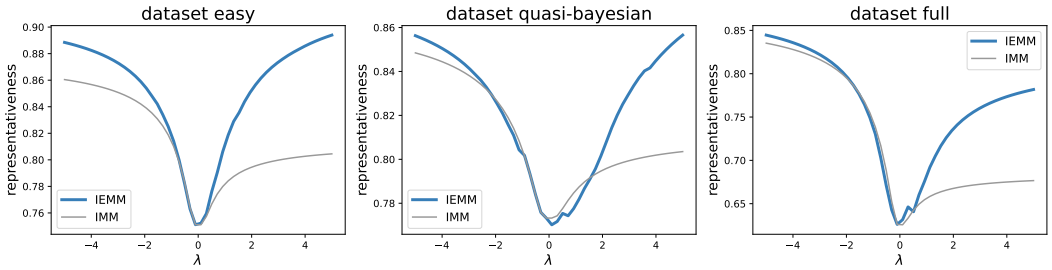

Figure 9: Overview of evidential representativeness across utilities and λ on synthetic datasets ($\mathcal{M}_{\text{easy}}$, $\mathcal{M}_{\text{full}}$, and $\mathcal{M}_{\text{quasi-bayesian}}$). The λ-evidential mistakeness typically yields the best $\mathcal{U}^{\lambda}$-representativeness for its corresponding utility.

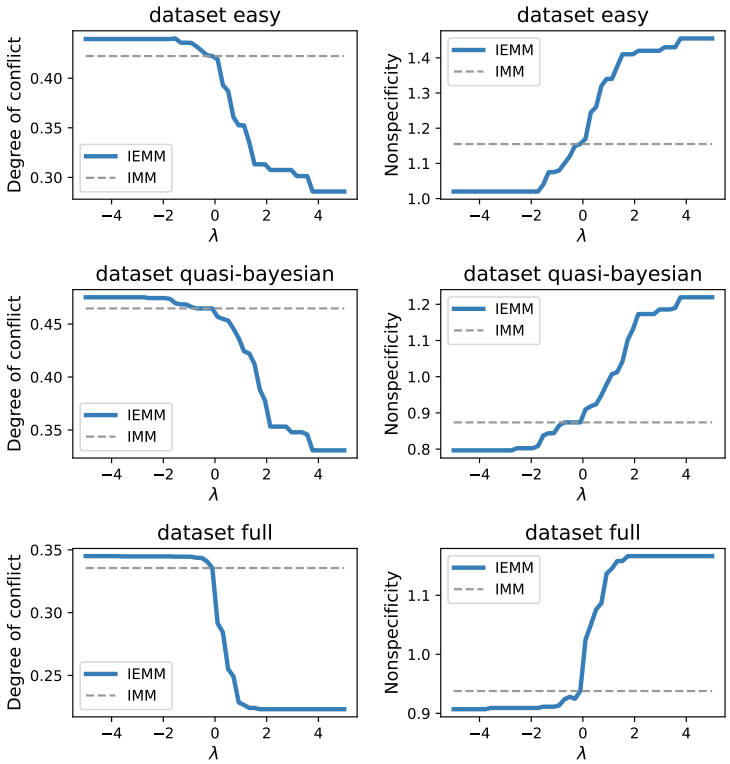

Figure 10: Conflict/non-specificity analysis for the synthetic experiments across λ and utilities. As λ increases, explanations become more cautious (lower conflict, higher non-specificity), in line with the trade-off in Denoux et al. (2018).

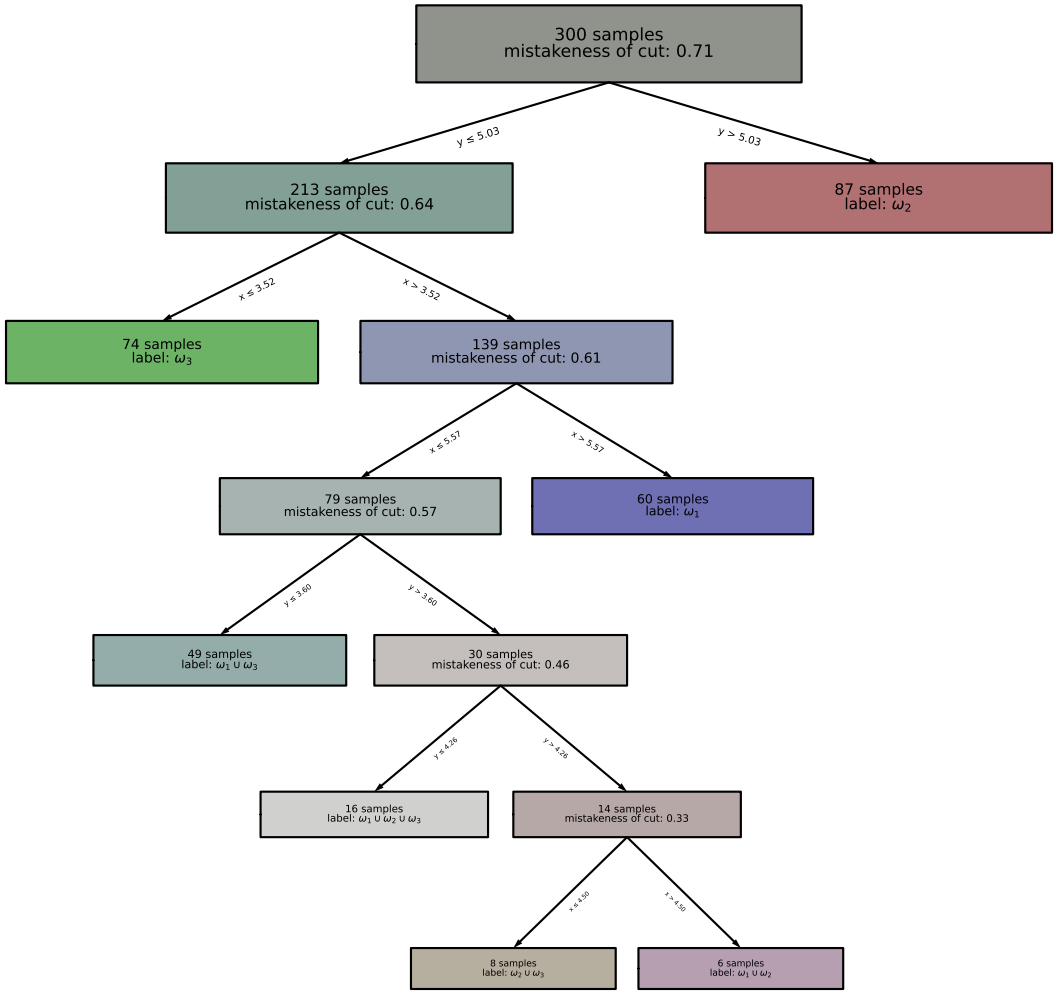

Figure 11: Decision tree obtained with IEMM for the evidential clustering function $\mathcal{M}_{\text{full}}$ and $\lambda = 0$. The decision trees generated by IEMM are shallow by construction, having at most $|\mathbb{F}| - 1$ levels. Each non-terminal node indicates the mistakeness of the corresponding split.

	ω_2	$\omega_1 \cup \omega_2 \cup \omega_3$	ω_3	ω_1
$M^{-\infty}_{\mathcal{M}_{q-bay}}$	$(y \leq 4.54) \wedge (x \leq 4.43)$	$(y \leq 4.54) \wedge (x > 4.43) \wedge (x \leq 4.48)$	$(y \leq 4.54) \wedge (x > 4.48)$	$(y > 4.54)$
$M^{-1}_{\mathcal{M}_{q-bay}}$	$(y \leq 4.54) \wedge (x \leq 4.08)$	$(y \leq 4.54) \wedge (x > 4.08) \wedge (x \leq 4.98)$	$(y \leq 4.54) \wedge (x > 4.98)$	$(y > 4.54)$
$M^{0}_{\mathcal{M}_{q-bay}}$	$(y \leq 4.69) \wedge (x \leq 3.85)$	$(y \leq 4.69) \wedge (x > 3.85) \wedge (x \leq 5.09)$	$(y \leq 4.69) \wedge (x > 5.09)$	$(y > 4.69)$
$M^{1}_{\mathcal{M}_{q-bay}}$	$(y \leq 5.39) \wedge (x \leq 3.68)$	$(y \leq 5.39) \wedge (x \leq 5.28) \wedge (x > 3.68)$	$(y \leq 5.39) \wedge (x > 5.28)$	$(y > 5.39)$
$M^{\infty}_{\mathcal{M}_{q-bay}}$	$(y \leq 5.82) \wedge (x \leq 2.95)$	$(y \leq 5.82) \wedge (x \leq 5.95) \wedge (x > 2.95)$	$(y \leq 5.82) \wedge (x > 5.95)$	$(y > 5.82)$

Table 2: Abductive explanations generated by IEMM for all clusters of the quasi-bayesian evidential clustering function. Higher λ values result in larger portions of the feature space being attributed to the cautious metacluster $\omega_1 \cup \omega_2 \cup \omega_3$.

	$\mathcal{R}_{\mathcal{M}_{easy},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{easy},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{easy},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{easy},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{easy},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{easy}}$	**0.915796**	0.808588	0.701380	0.701452	0.701524
$M^{-1}_{\mathcal{M}_{easy}}$	0.901122	**0.819012**	0.736903	0.749377	0.761850
$M^{0}_{\mathcal{M}_{easy}}$	0.876733	0.813867	**0.751002**	0.781731	0.812461
$M^{1}_{\mathcal{M}_{easy}}$	0.781247	0.751198	0.721149	**0.811562**	0.901975
$M^{\infty}_{\mathcal{M}_{easy}}$	0.689432	0.669249	0.649067	0.789613	**0.930160**
	$\mathcal{R}_{\mathcal{M}_{full},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{full},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{full},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{full},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{full},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{full}}$	**0.882009**	0.731303	0.575128	0.593766	0.612354
$M^{-1}_{\mathcal{M}_{full}}$	0.881689	0.738726	0.598038	0.618498	0.638218
$M^{0}_{\mathcal{M}_{full}}$	0.867413	**0.745508**	**0.625343**	0.656206	0.683719
$M^{1}_{\mathcal{M}_{full}}$	0.642447	0.596290	0.542490	**0.681838**	0.809441
$M^{\infty}_{\mathcal{M}_{full}}$	0.621851	0.577133	0.526137	0.679054	**0.818054**
	$\mathcal{R}_{\mathcal{M}_{q-bay},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{q-bay},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{q-bay},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{q-bay},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{q-bay},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{q-bay}}$	**0.887444**	0.781227	0.728118	0.728120	0.728125
$M^{-1}_{\mathcal{M}_{q-bay}}$	0.872286	0.799687	0.763387	0.772136	0.789634
$M^{0}_{\mathcal{M}_{q-bay}}$	0.866938	**0.804483**	**0.773256**	**0.785821**	0.810953
$M^{1}_{\mathcal{M}_{q-bay}}$	0.802770	0.761888	0.741446	0.778781	0.853452
$M^{\infty}_{\mathcal{M}_{q-bay}}$	0.638852	0.617576	0.606938	0.708914	**0.912866**

Table 3: Evaluation of the resulting explanations for each metacluster of the synthetic datasets. Each line corresponds to a decision tree trained with one specific mistakeness. Each column corresponds to the $\mathcal{U}$-evidential representativeness of the decision tree. In bold, the best decision tree for each representativeness. We can see that decision trees trained with λ-evidential mistakeness function tend to be the best in terms of $\mathcal{U}^{\lambda}$-evidential representativeness.

G.2 Testing IEMM as a Classifier on Real-World Datasets

To further validate our approach, we also assess IEMM as a stand-alone evidential classifier on larger datasets from the `credal-datasets-master` repository Hoarau et al. (2023b). Unlike explaining a given clustering, this setting requires fitting an evidential model directly from features—learning both decision boundaries and mass assignments—making the task more challenging. We report representativeness in Table 4, provide an overview of per-λ behavior in Figure 12, and analyze the conflict/non-specificity trade-off in Figure 13.

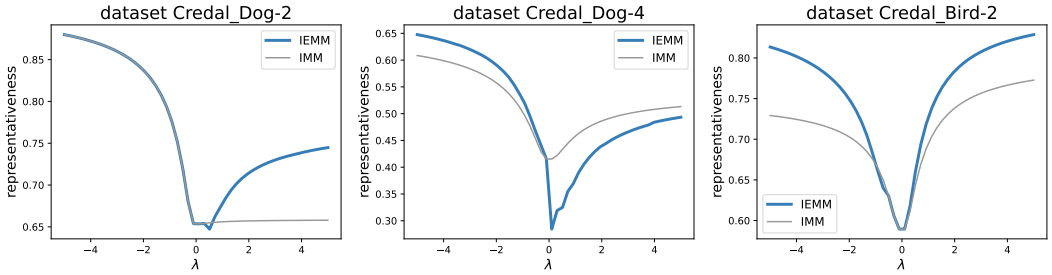

Figure 12: Overview of IEMM classification results across utilities and λ on credal datasets. Higher λ values lead to more cautious decisions, while lower values favor specificity.

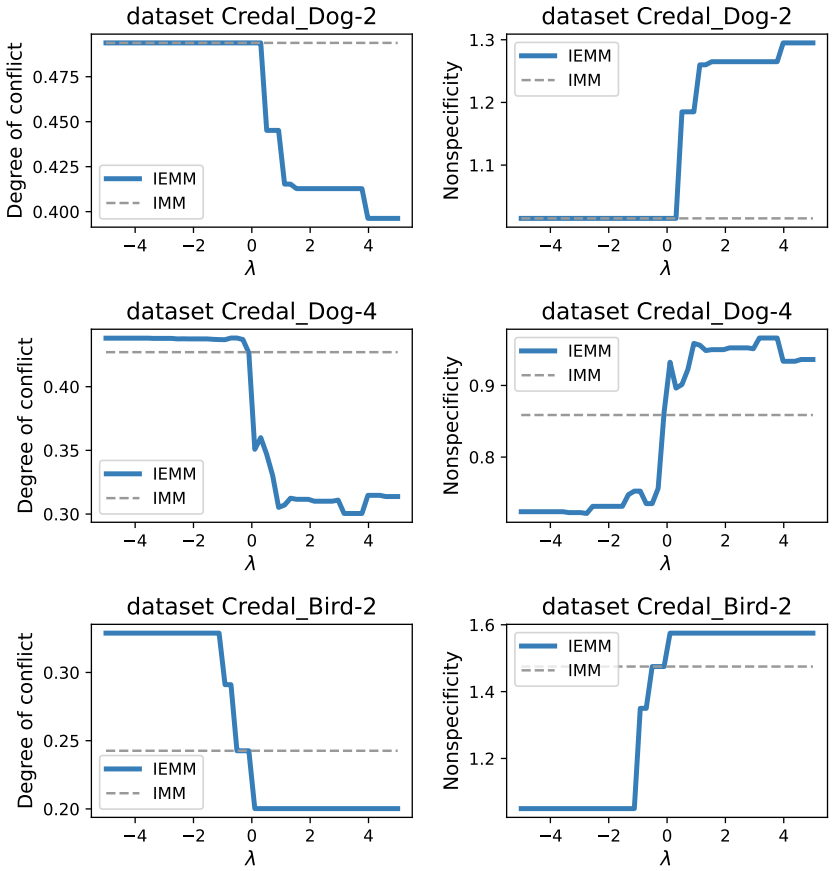

Figure 13: Conflict and non-specificity for IEMM classification across λ on credal datasets. The λ knob shifts the operating point along the conflict/non-specificity frontier, echoing the behavior observed in clustering.

G.3 Extra Results on Explaining the Clustering in Real-World Datasets

Table 5 reports the corresponding representativeness scores on Iris, Wine, and Diabetes.

	$\mathcal{R}_{\mathcal{M}_{CB-2},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{CB-2},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{CB-2},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{CB-2},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{CB-2},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{CB-2}}$	**0.864286**	0.658929	0.453571	0.458929	0.464286
$M^{-1}_{\mathcal{M}_{CB-2}}$	**0.864286**	0.667857	0.471429	0.480357	0.489286
$M^{0}_{\mathcal{M}_{CB-2}}$	0.750000	**0.669643**	**0.589286**	0.694643	0.800000
$M^{1}_{\mathcal{M}_{CB-2}}$	0.714286	0.651786	**0.589286**	**0.726786**	**0.864286**
$M^{\infty}_{\mathcal{M}_{CB-2}}$	0.714286	0.651786	**0.589286**	**0.726786**	**0.864286**

	$\mathcal{R}_{\mathcal{M}_{CD-2},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{CD-2},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{CD-2},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{CD-2},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{CD-2},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{CD-2}}$	**0.913571**	0.780357	0.647143	0.647857	0.648571
$M^{-1}_{\mathcal{M}_{CD-2}}$	**0.913571**	**0.783571**	**0.653571**	0.656071	0.658571
$M^{0}_{\mathcal{M}_{CD-2}}$	**0.913571**	**0.783571**	**0.653571**	0.656071	0.658571
$M^{1}_{\mathcal{M}_{CD-2}}$	0.775714	0.682500	0.589286	**0.677500**	0.765714
$M^{\infty}_{\mathcal{M}_{CD-2}}$	0.661429	0.575000	0.488571	0.632500	**0.776429**

	$\mathcal{R}_{\mathcal{M}_{CD-4},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{CD-4},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{CD-4},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{CD-4},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{CD-4},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{CD-4}}$	**0.696531**	0.523750	0.416378	0.441318	0.470153
$M^{-1}_{\mathcal{M}_{CD-4}}$	0.694184	**0.526565**	**0.422959**	0.454137	0.498980
$M^{0}_{\mathcal{M}_{CD-4}}$	0.653622	0.503031	0.414898	**0.458236**	0.537194
$M^{1}_{\mathcal{M}_{CD-4}}$	0.467959	0.345948	0.242449	0.378104	0.543163
$M^{\infty}_{\mathcal{M}_{CD-4}}$	0.421224	0.320476	0.236990	0.388116	**0.601071**

Table 4: Evaluation of the resulting explanations for each metacluster of the real-world datasets. Each line corresponds to a decision tree trained with one specific mistakeness. We implemented the IEMM algorithm over the datasets `Credal_Bird-2` ($\mathcal{M}_{CB-2}$ with 2 classes), `Credal_Dog-2` ($\mathcal{M}_{CD-2}$ with 2 classes) and `Credal_Dog-4` ($\mathcal{M}_{CD-4}$ with 4 classes). Each column corresponds to the $\mathcal{U}$-evidential representativeness of the decision tree. In bold, the best decision tree for each representativeness. We can see that decision trees trained with λ-evidential mistakeness function tend to be the best in terms of $\mathcal{U}^{\lambda}$-evidential representativeness.

	$\mathcal{R}_{\mathcal{M}_{iris},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{iris},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{iris},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{iris},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{iris},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{iris}}$	**0.872336**	0.746751	0.619907	0.633296	0.646797
$M^{-1}_{\mathcal{M}_{iris}}$	0.847368	0.776900	0.707914	0.742323	0.777439
$M^{0}_{\mathcal{M}_{iris}}$	0.846415	**0.778853**	**0.712790**	0.745366	0.777935
$M^{1}_{\mathcal{M}_{iris}}$	0.773996	0.725864	0.671799	**0.749179**	**0.838243**
$M^{\infty}_{\mathcal{M}_{iris}}$	0.647244	0.595757	0.538105	0.678586	0.830750

	$\mathcal{R}_{\mathcal{M}_{wine},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{wine},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{wine},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{wine},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{wine},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{wine}}$	**0.999079**	0.906952	0.814824	0.814824	0.814824
$M^{-1}_{\mathcal{M}_{wine}}$	0.994454	**0.973001**	0.951548	0.953907	0.956266
$M^{0}_{\mathcal{M}_{wine}}$	0.986110	0.971693	**0.957276**	0.963835	0.970394
$M^{1}_{\mathcal{M}_{wine}}$	0.971606	0.963952	0.956299	**0.970139**	0.983980
$M^{\infty}_{\mathcal{M}_{wine}}$	0.623439	0.621431	0.619423	0.807394	**0.995364**

	$\mathcal{R}_{\mathcal{M}_{diabetes},\mathcal{U}^{-\infty}}$	$\mathcal{R}_{\mathcal{M}_{diabetes},\mathcal{U}^{-1}}$	$\mathcal{R}_{\mathcal{M}_{diabetes},\mathcal{U}^{0}}$	$\mathcal{R}_{\mathcal{M}_{diabetes},\mathcal{U}^{1}}$	$\mathcal{R}_{\mathcal{M}_{diabetes},\mathcal{U}^{\infty}}$
$M^{-\infty}_{\mathcal{M}_{diabetes}}$	**0.854390**	0.698233	0.542076	0.542663	0.543251
$M^{-1}_{\mathcal{M}_{diabetes}}$	0.816186	**0.701698**	0.587211	0.642283	0.697355
$M^{0}_{\mathcal{M}_{diabetes}}$	0.783105	0.689771	**0.596438**	0.675605	0.754772
$M^{1}_{\mathcal{M}_{diabetes}}$	0.767454	0.680228	0.593003	**0.681899**	0.770795
$M^{\infty}_{\mathcal{M}_{diabetes}}$	0.687602	0.614695	0.541787	0.674057	**0.806328**

Table 5: Representativeness of IEMM explanations on Iris, Wine, and Diabetes across λ values. Higher is better. Best scores per representativeness are highlighted in the tables.