Credal discrete classifier

Wenlong Chen[1] **Cyprien Gilet**[1] **Benjamin Quost**[1] **Sébastien Destercke**[1]

[1]UMR CNRS 7253 Heudiasyc, Université de Technologie de Compiègne, France

Abstract

This paper presents a novel extension of the discrete Bayesian classifier (DBC) into a set-valued classification framework grounded in imprecise probability theory. The standard DBC framework, which relies on partitioning the input space into profiles and estimating class-conditional probabilities, may not be very robust to distribution changes or imperfections in observed data. In the hope to mitigate such issues, we introduce the Credal Discrete Classifier (CDC), an imprecise-probabilistic extension of the traditional Bayesian approach. By representing uncertainties in the estimated probabilities through belief functions, CDC offers interval-valued risks and set-valued decisions, thereby enhancing robustness. Experimental results on several benchmark datasets demonstrate that CDC effectively balances accuracy and determinacy by allowing for set-valued predictions in uncertain contexts, often outperforming or matching traditional precise classifiers.

Keywords. Bayesian classifier, credal classifier, belief function, decision criteria

1. Introduction

Classification tasks consist in assigning labels to instances; this is generally done so as to minimize an expected cost or loss. Standard methods like the Bayesian classifier achieve this by modeling the underlying probability distribution and selecting the label that minimizes the expected misclassification cost. However, this approach typically assumes that the training distribution is representative of future, real-world data, and that observed data are sufficient in quantity and quality. This may not hold, due to the available training data being scarce, imperfect, or due to a shift between the training and test distributions: in many practical scenarios, such as medical diagnosis, financial forecasting, or fault detection in industrial processes, the data-generating process may indeed evolve over time or differ between domains.

Discrete Bayesian classifiers (DBC) provide a flexible nonparametric means of estimating posterior probabilities by partitioning the input space into *profiles*. Although appealing for data with complex structures, DBC predictions can still be misleading when distributional assumptions fail to hold—such as under covariate shift or in presence of limited data. To address this shortcoming, Soft Probabilistic Discrete Bayesian classifiers (SPDBC) leverage a soft partitioning of the input space, making the decision boundaries less dependent to local information (pertaining to a single profile), and thereby arguably improving generalization.

Despite these advancements, handling a single posterior probability distribution estimate still makes the model vulnerable to shift or limited training data. We propose a robustified version of the SPDBC, the purpose of which is to ultimately overcome these issues. In particular, we replace the fixed posterior probability with a set of plausible distributions, thereby yielding interval-valued risks. We coin the resulting method Credal Discrete Classifier (CDC). By leveraging the lower and upper bounds on misclassification risk, CDC can be combined with different decision criteria—such as interval dominance, or component-wise strict dominance—to produce either precise or set-valued decisions. These set-valued outputs are particularly beneficial in safety-critical applications, where a robust, partially uncertain prediction may be preferable to an overconfident, single-label decision.

This article has the following structure. Section 2 first reviews the background and definition of DBC and SPDBC. Section 3 formally derives the CDC by using probability sets, and discusses several decision criteria. Section 4 proposes an evidential instantiation of the CDC and approximations to improve computational efficiency. Section 5 reports experiments where we compare the performances between the CDC and SPDBC models. Section 6 concludes the paper.

2. Discrete Bayesian classifier

2.1. Problem setting. We seek to construct a classification function $\delta : \mathcal{X} \to \mathcal{Y}$ that predicts, for any input instance $x \in \mathcal{X}$, the corresponding class label $\delta(x) \in \mathcal{Y} = \{1, \dots, K\}$. This function is learned from a training set $\{(x_1, y_1), \dots, (x_n, y_n)\}$ of observed instance-label pairs, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ represent the feature vector and true class label of the i-th instance, respectively.

The problem is formulated under asymmetric deci-

sion costs, where misclassification penalties may differ substantially between error types. This cost structure is encoded in a non-negative matrix L_{kl}, where $L_{kl} \geq 0$ denotes the penalty for predicting class l when the true class is k. While diagonal entries satisfy $L_{kk} = 0$ (correct classifications incur no cost), off-diagonal entries may be asymmetric ($L_{kl} \neq L_{lk}$ for $k \neq l$). For instance, in medical diagnosis, false positives typically carry lower penalties than false negatives.

2.2. Bayes' decision strategy. Bayesian decision theory [2, 6, 18] provides a theoretical solution to this learning problem. Under the assumption that data are generated according to a joint probability $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$, it establishes that the decision strategy minimizing the *expected risk* (or misclassification loss) should be based on the posterior probabilities of the classes and on the misclassification costs:

$$\delta^B(x) = \arg\min_{l \in \mathcal{Y}} R_l(x), \tag{1a}$$

$$\text{where} \quad R_l(x) = \sum_{k \in \mathcal{Y}} L_{kl}\, \mathbb{P}(Y = k \mid X = x). \tag{1b}$$

Generative models typically derive the posterior probabilities $\mathbb{P}(Y = k|X = x)$ from the class-conditional distributions $\mathbb{P}(X = x|Y = k)$ and the prior probabilities π_k using Bayes' rule:

$$\mathbb{P}(Y = k \mid X = x) = \frac{\pi_k \mathbb{P}(X = x \mid Y = k)}{\mathbb{P}(X = x)},$$

$$\text{with} \quad \mathbb{P}(X = x) = \sum_{l=1}^{K} \pi_l \mathbb{P}(X = x \mid Y = l).$$

Therefore, a wide range of approaches aim at estimating the prior probabilities π_k and the conditional distributions $\mathbb{P}(X = x|Y = k)$ so as to implement the Bayes classifier [1, 9, 10], for instance via maximum likelihood (ML). While π_k can be estimated using the class frequencies in the training set, estimating $\mathbb{P}(X = x|Y = k)$ usually requires additional assumptions. Thus, many strategies postulate a (semi-)parametric model for $\mathbb{P}(X = x|Y = k)$ and focus on estimating the parameters of the distribution, for instance as in discriminant analysis [11].

2.3. Discrete Bayesian classifier. When the assumed distributional model does not hold, the resulting classifier may become a biased estimate of the true Bayes classifier [10], leading to suboptimal performance even with large training samples. A viable alternative is to adopt a nonparametric approach. The Discrete Bayesian Classifier (DBC) achieves this by partitioning the input space $\mathcal{X}$ into a set of T discrete profiles, denoted as $\mathcal{T} = \{\phi_1, \ldots, \phi_T\}$. A mapping function $\Phi : \mathcal{X} \to \mathcal{T}$ is defined to assign each instance x to a specific profile ϕ_t. We use the notation $\Phi(x) = \phi_t$ or equivalently $x \in \phi_t$ to indicate that instance x belongs to profile ϕ_t.

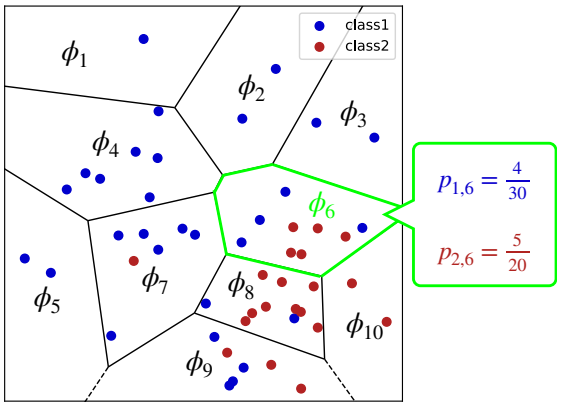

Figure 1. *Bivariate classification problem using DBC. Class 1 (blue) counts $n_1 = 30$ training instances, and class 2 (red) $n_2 = 20$ training instances.*

Instead of estimating the full class-conditional density $\mathbb{P}(X = x|Y = k)$, the DBC approximates it by computing the probability of an instance belonging to each profile given its class label:

$$\hat{p}_{kt} = \hat{\mathbb{P}}(\Phi(X) = \phi_t \mid Y = k) := \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{\Phi(x_i) = \phi_t\}}, \tag{2}$$

where $\mathcal{I}_k = \{i \in \{1, \ldots, n\} \mid Y_i = k\}$ is the index set of training instances belonging to class k, $n_k = |\mathcal{I}_k|$ is the number of training samples in class k, and $\mathbb{1}_{\{\cdot\}}$ is an indicator function allowing to check for profile membership:

$$\mathbb{1}_{\{\Phi(x) = \phi_t\}} = \begin{cases} 1 & \text{if } x \text{ is assigned to profile } \phi_t, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

This approach introduces two levels of approximation: the *discretization step*, where the continuous input space is divided into profiles, effectively replacing $\mathbb{P}(X = x|Y = k)$ with $\mathbb{P}(\Phi(X) = \phi_t \mid Y = k)$; and the *estimation step*, where the latter probability is approximated using empirical class frequencies within each profile. Figure 1 provides a visual example of DBC applied to a two-dimensional dataset, where K-means is used for profile partitioning.

Given these profile-based probability estimates, the DBC classifier assigns each instance x to the class that minimizes the expected risk [8]:

$$R_l(x) := \sum_{k \in \mathcal{Y}} L_{kl} \sum_{t=1}^{T} \pi_k \hat{p}_{kt} \mathbb{1}_{\{\Phi(x) = \phi_t\}}, \quad \forall l \in \mathcal{Y}. \tag{4}$$

Here, the risk $R_l(x)$ is computed as a weighted sum of misclassification costs across all profiles, taking into account the prior class probabilities π_k. This formulation makes DBC a simple yet flexible classifier, particularly

useful when the underlying class-conditional distributions are complex [7, 8] and call for a nonparametric modeling.

2.4. Soft probabilistic discrete Bayesian classifier. Traditional DBC typically employs a hard partitioning scheme, which induces several challenges [4]. First, the risk function $R_l(x)$ remains identical for all instances within the same profile, causing decision boundaries to align strictly with profile boundaries. This dependence makes the classifier performance highly sensitive to the choice of the partitioning algorithm. Consequently, two nearby instances may be assigned different labels due to being into different profiles. In addition, instances close to x in the input space but assigned to different profiles have no influence on the estimated class-conditional risks in Equation (4). As a result, the decision boundary becomes a union of multiple locally optimal, profile-based partitions, rather than a globally optimal solution.

To mitigate these issues, we introduced in [4] the Soft Probabilistic Discrete Bayesian Classifier (SPDBC), which replaces hard profile assignments with soft cluster memberships. This allows each instance to belong to multiple profiles with different degrees of confidence:

$$R_l(x) := \sum_{k\in\mathcal{Y}} L_{kl} \sum_{t=1}^{T} \pi_k \widehat{p}_{kt} p_t^*, \quad \forall l \in \mathcal{Y}, \tag{5a}$$

$$\widehat{p}_{kt} := \frac{1}{n_k} \sum_{i\in\mathcal{I}_k} p_t^{(i)}. \tag{5b}$$

The probabilities $p_t^{(i)} = \mathbb{P}(\Phi(x_i) = \phi_t)$ and $p_t^* = \mathbb{P}(\Phi(x) = \phi_t)$ can be obtained from various soft clustering techniques, such as fuzzy c-means [3] or Gaussian mixture models [5]. Here, the superscript (i) denotes the i-th instance in the training set, while the superscript $*$ refers to the new instance to be classified. By allowing instances to belong to multiple profiles with different degrees, SPDBC introduces a more flexible decision boundary, which significantly improves the generalization performance and robustness to variations in the profile structure. For notational simplicity, we will refer by $p_t = \mathbb{P}(\Phi(X) = \phi_t)$ to the probability of the profile membership ϕ_t when X is treated as a generic random variable. Both $p_t^{(i)}$ and p_t^* can be viewed as being derived from this general quantity.

3. Credal discrete classifier

3.1. Motivation: from probabilities to credal sets. Traditional classifiers, including the Discrete Bayesian Classifier (DBC) and its soft extension (SPDBC), rely on estimating class-conditional probabilities based on training data. However, these approaches assume that the estimated probabilities remain valid in future scenarios. In practice, this assumption often fails due to distribution shifts, where the test data distribution differs from the training distribution.

A key limitation of DBC and SPDBC is that they provide point estimates of probabilities, assuming a single, fixed probability distribution for classification. This assumption may not hold in real-world applications, for various reasons. Indeed, the feature distribution $\mathbb{P}(X)$ may change over time (covariate shift), even if the conditional distribution $\mathbb{P}(Y|X)$ remains stable. As well, when data is scarce, the estimated class-conditional probabilities may be highly uncertain and unreliable. For example, if only a few training samples are available for a given profile, the empirical estimate of $\mathbb{P}(Y|X)$ can be dominated by noise or sampling variability, leading to overconfident or misleading predictions. This is particularly problematic for profiles with rare feature combinations. Last, the profile probabilities $p_t^{(i)}$ and p_t^* in the SPDBC are estimated from data, but the profile assignment of new test samples may vary significantly from those of the training samples.

To address these issues, we propose to move from precise probabilities to *credal sets* [13, 22], which represent a family of plausible probability distributions instead of a single fixed estimate. This allows us to model uncertainty explicitly, and to make decisions that account for worst-case and best-case scenarios.

In the DBC and SPDBC, the class-conditional probabilities $\widehat{p}_{kt} = \widehat{\mathbb{P}}(\Phi(X) = \phi_t \mid Y = k)$ are estimated directly from the training data, and the profile assignment probabilities $p_t = \mathbb{P}(\Phi(X) = \phi_t)$ are obtained via a clustering algorithm. To model uncertainty, replace the single estimated probability distribution with a set of plausible distributions:

$$p_t \in \mathcal{P}_t,$$

where each $\mathcal{P}_t$ represents a credal set, *i.e.*, a collection of possible distributions that account for estimation errors. By adopting this framework, we can compute intervals instead of point estimates, leading to a more robust decision-making process in the presence of uncertainty.

3.2. Credal set computation. Moving from precise probability estimates to sets of distributions raises the question of how to specify or characterize the credal sets $\mathcal{P}_t$ for each profile ϕ_t. In practice, these constraints may be derived from various principled approaches that capture different types of uncertainty. Below, we briefly discuss three illustrative methods for eliciting credal constraints in discrete classification: evidential clustering, the imprecise Dirichlet model (IDM), and neighborhood-based constraints.

Evidential clustering. When profiles $\{\phi_1, \dots, \phi_T\}$ are obtained through an evidential clustering algorithm (e.g., ECM [14]), each instance x is assigned a mass function $m(\Phi(x) = A)$ over subsets $A \subseteq \{\phi_1, \dots, \phi_T\}$. These mass functions naturally induce lower and upper probabili-

ties (belief and plausibility) for each profile [19], thus defining

$$\mathrm{Bel}(\Phi(x) = \phi_t) \;\leq\; p_t \;\leq\; \mathrm{Pl}(\Phi(x) = \phi_t). \qquad (6)$$

Aggregating such credal constraints across training samples for each class k leads to intervals $\widehat{\mathrm{Bel}}_{kt}$ and $\widehat{\mathrm{Pl}}_{kt}$ (respectively the empirical belief and plausibility) in place of single-value estimates $\widehat{p}_{kt}$. Similarly, for a new instance x, evidential clustering yields a mass function whose induced lower and upper probabilities constrain p_t^*. This approach, detailed in Section 4, can soften decision bounds and potentially protect our method from data noise.

Imprecise Dirichlet Model (IDM). Another classical way to build credal sets for multinomial parameters is the IDM [22] and its posterior inferences.Concretely, if α_{kt} denotes the prior imprecision on profile ϕ_t under class k, the posterior $\widehat{p}_{kt}$ lies within an interval derived from

$$\widehat{p}_{kt} \;\in\; \Big[\frac{n_{kt}}{n_k + \sum_u \alpha_{ku}}, \frac{n_{kt} + \alpha_{kt}}{n_k + \sum_u \alpha_{ku}}\Big],$$

where n_{kt} is the empirical count of samples from class k assigned to profile ϕ_t, and n_k is the total number of samples of class k. This approach is particularly appealing when the training sample size is small or highly unbalanced, thus offering a possible solution to such imperfections.

Neighborhood-based constraints. A simple way to make p_t imprecise is to consider its initial estimates $\hat{p}_t$ and to consider a neighbourhoud $\mathcal{P}_t$ centered around it. There are many ways to induce such neighbourhouds [16], and those could be used to describe a possible drift of $\hat{p}_t$ within the space $\mathcal{P}_t$.

Imprecise data constraints. Assume that a data point x_i is not precisely observed, but is only known to belong to a set X_i. In this case, the estimator of $\hat{p}_{kt}$ given by Equation 2 is no longer precisely defined, as X_i could belong to multiple profiles at once. This would induce some imprecision in $\hat{p}_{kt}$, due here to the imprecision of observations.

In the present work, we rely on evidential clustering methods of type (1) to define the credal constraints $\mathcal{P}_t$. In particular, we use Evidential C-Means (ECM) [14] as a natural extension of the fuzzy C-means algorithm which used in SPDBC. ECM inherits the intuitive semantics of soft clustering while additionally enabling the expression of epistemic uncertainty through mass functions. This choice ensures both interpretability and computational tractability. Nevertheless, the alternative strategies discussed above, especially those based on local neighborhood structure or Bayesian imprecise modeling (e.g., IDM), offer promising directions for future work.

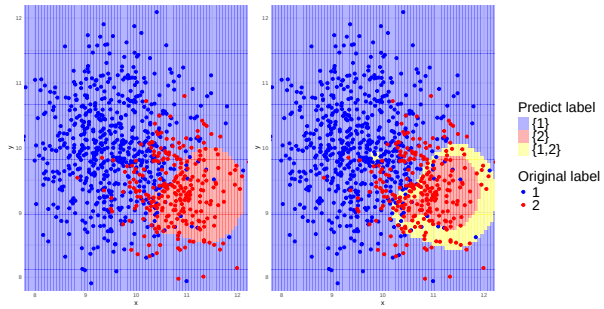

Figure 2. *Bivariate classification problem with decision boundaries of SPDBC (left) and CDC (right).*

3.3. Formulation of the credal discrete classifier. Building upon the concept of credal sets, the risk function becomes:

$$R_l(x) := \sum_{k \in y} L_{kl} \sum_{t=1}^{T} \pi_k \, \widehat{p}_{kt} \, p_t^*, \quad \forall l \in \mathcal{Y}, \qquad (7a)$$

$$\text{s.t.} \quad p_t \in \mathcal{P}_t. \qquad (7b)$$

Since the probability values p_t are no longer fixed but instead vary within a credal set $\mathcal{P}_t$, the resulting risk $R_l(x)$, which depends linearly on these probabilities, also becomes a variable quantity rather than a single scalar. To accommodate for this imprecision and prepare for decision-making under uncertainty, we first characterize the range of possible risk values. Specifically, we consider the interval-valued risk defined by:

$$R_l(x) \in \Big[\underline{\mathcal{R}_l}(x), \overline{\mathcal{R}_l}(x)\Big]. \qquad (8)$$

Here, $\underline{\mathcal{R}_l}(x)$ and $\overline{\mathcal{R}_l}(x)$ respectively represent the lowest (optimistic assumption) and highest (pessimistic assumption) possible risks. As a consequence, the traditional approach consisting in minimizing the risk is no longer directly applicable, necessitating alternative decision-making strategies to make robust predictions in uncertain environments.

Finding the $R_l(x)$ can be defined by the following nonlinear optimization problem:

$$\min / \max R_l(x) \quad \text{s.t.} \quad p_t \in \mathcal{P}_t.$$

Under these conditions, our model functions as a credal classifier [24]. Depending on the selected decision strategy, the classifier's output can be a set-valued prediction, such as those obtained through interval dominance or maximality, or a single-label prediction, such as those derived using maximin or maximax criteria.

3.4. Decision-making criteria. Once we obtain the risk interval as per Equation (8), different decision criteria can be applied to determine the classification outcome. In our model, we employ *maximality*, *interval dominance*

and *component-wise strict dominance* [21], allowing the model to output multiple classes when the data is ambiguous or uncertain.

Given two classes n and m, for any possible distribution, if $R_n(x)$ is smaller smaller then $R_m(x)$, then class n dominates class m according to maximality if

$$R_m(x) - R_n(x) > 0, \quad \forall p_t \in \mathcal{P}_t.$$

This can be checked by solving the following optimization problem:

$$\min \sum_{k \in \mathcal{Y}} (L_{km} - L_{kn}) \sum_{t=1}^{T} \pi_k \widehat{p}_{kt} p_t^* > 0, \quad \forall m, n \in \mathcal{Y},$$
$$\text{s.t.} \quad p_t \in \mathcal{P}_t.$$

We retain all classes that are not dominated by any other class.

The *interval dominance* strategy is based on the following principle: given two classes n and m, if the maximum risk of class n is strictly smaller than the minimum risk of class m, *i.e*,

$$\overline{\mathcal{R}_n}(x) < \underline{\mathcal{R}_m}(x),$$

then class n interval-dominates class m. Ultimately, we retain all classes that are not interval-dominated by any other class.

The *component-wise strict dominance* strategy, as defined in this work, considers the two components of a risk interval—its lower and upper bounds—as dimensions for comparison. Given two classes n and m, if both the minimum risk and maximum risk of class n are strictly smaller than those of class m, *i.e*.,

$$\underline{\mathcal{R}_n}(x) < \underline{\mathcal{R}_m}(x) \quad \text{and} \quad \overline{\mathcal{R}_n}(x) < \overline{\mathcal{R}_m}(x),$$

then class n is said to component-wise strictly dominate class m. This criterion ensures that n is safer than m in both the best- and worst-case scenarios.

Both *interval dominance* and *component-wise strict dominance* utilize the maximum and minimum risk, ensuring that both the worst-case scenario (pessimistic decision-making) and the best-case scenario (optimistic decision-making) are taken into account. In contrast, *maximality* adopts a different perspective: instead of relying directly on risk bounds, it considers the dominance of one class over another with respect to all admissible probability distributions. Due to this property, maximality does not explicitly involve the minimum or maximum risks. In Section 4.3, we come back to this point and elaborate on how maximality can be used in practice.

Note that interval dominance is more conservative than maximality, itself more conservative than component-wise strict dominance: this notably implies that the subsets of classes obtained include one another. However, this only holds if these subsets are exactly computed. Since for computational reasons we consider approximations of those criteria, it seems useful to check whether this inclusion still holds empirically. This will indeed be the case in our experiment.

4. Instanciation

To effectively estimate the risk bounds in the presence of uncertainty, we employ an evidential framework based on Dempster–Shafer theory. Rather than relying on a single probability value, we represent the uncertainty in profile assignments and conditional probabilities using belief functions, which define lower and upper probabilities for each hypothesis.

In our approach, we consider the set of profiles $\{\phi_1, ..., \phi_T\}$ as a frame of discernment, denoted by Ω, and assign mass functions $m(\Phi(X) = A)$ to subsets $A \subseteq \Omega$. These mass functions quantify the degree of belief that an instance belongs to a particular subset of profiles. Given a profile assignment function Φ, the belief and plausibility of an instance belonging to a specific profile ϕ_t are defined as

$$\text{Bel}_t = \text{Bel}(\Phi(X) = \phi_t) = \sum_{A \subseteq \Omega, A \subseteq \phi_t} m(A), \tag{9}$$

$$\text{Pl}_t = \text{Pl}(\Phi(X) = \phi_t) = \sum_{A \subseteq \Omega, A \cap \phi_t \neq \emptyset} m(A). \tag{10}$$

The former represents the minimum probability mass that supports the hypothesis that X belongs to ϕ_t, the latter provides an upper bound on the probability that X belongs to ϕ_t by including all subsets A that intersect ϕ_t.

The mass function can be provided by any evidential clustering algorithm (e.g., Evidential C-means or Evidential K-NN), which compute a credal partition of the data, *i.e*. which assign each $x \in \mathcal{X}$ to (subsets of) profiles A with specific mass degrees. As a matter of fact, m implicitly defines a family of distributions (a credal set) $\mathcal{P}_t = \mathcal{P}(\Phi(X) = \phi_t)$:

$$\sum_{t=1}^{T} p_t = 1, \quad \text{Bel}_t \leq p_t \leq \text{Pl}_t. \tag{11}$$

The challenge is to compute the lower and upper bounds on $R_l(x)$ under the constraints (11) so as to use the decision criteria mentioned above. This can be done by solving the following bilinear optimization problem:

$$\min / \max R_l(x)$$
$$\text{s.t.} \quad \text{Bel}_t \leq p_t \leq \text{Pl}_t, \quad \sum_{t=1}^{T} p_t = 1.$$

4.1. Risk bound approximations. The complexity of this type of problem has been extensively studied in the literature [12, 17, 23] and is generally classified as NP-hard. Common approaches for obtaining approximate

solutions include convex relaxation methods, such as McCormick envelopes [15], and alternating optimization techniques. However, given that our model requires solving the problem for each prediction instance x and class label l, even employing McCormick envelopes results in high computational costs. For instance, on the IRIS dataset with 150 samples, a single experiment using McCormick envelopes and alternating optimization took over 40 seconds, highlighting the impracticality of global optimization for larger datasets.

To address this computational challenge, we propose a relaxation strategy that eliminates the simplex constraint —- *i.e.*, the requirement that all profile probabilities must sum to one—- while retaining only the interval constraint, thereby enabling a more efficient solution. Although this relaxation disregards the mutual exclusivity constraints between different profile probabilities—leading to a broader, potentially more conservative interval compared to the exact solution—it significantly reduces computational complexity. More importantly, this approximation still provides meaningful upper and lower risk estimates across various distribution shift scenarios, enhancing the adaptability and robustness of the discrete Bayesian classifier within an evidential framework. The effectiveness and efficiency of this approximation will be illustrated in the experimental section.

When retaining only the interval constraints, the problem simplifies considerably. It becomes evident that, when both the loss function L and prior probabilities π are strictly positive, the minimum and maximum values of the objective function $R_l(x)$ are attained when the decision variables take their respective extreme values. Specifically, the lower bound of $R_l(x)$ is achieved for $p_t = \mathrm{Bel}_t$, and the upper bound for $p_t = \mathrm{Pl}_t$.

By leveraging this simplification, we obtain a computationally feasible method that approximates risk bounds efficiently while maintaining robustness against distributional uncertainty.

4.2. Interpretation of relaxation on belief function. As mentioned above, relaxing the original optimization problem significantly alleviates the computational cost of computing the risk bounds. However, it remains essential to assess the validity of this relaxation. To do so, we explore its theoretical implications.

Under the proposed relaxation, the risk function is given by:

$$\begin{aligned}\underline{\mathcal{R}_l}(x) &= \sum_{k\in\mathcal{Y}} L_{kl} \sum_{t=1}^{T} \sum_{i\in\mathcal{I}_k} \frac{\pi_k}{n_k} \mathrm{Bel}_t^{(i)} \mathrm{Bel}_t^* \\ &= \sum_{k\in\mathcal{Y}} L_{kl} \sum_{t=1}^{T} \pi_k \widehat{\mathrm{Bel}}_{kt} \mathrm{Bel}_t^*, \\ \overline{\mathcal{R}_l}(x) &= \sum_{k\in\mathcal{Y}} L_{kl} \sum_{t=1}^{T} \sum_{i\in\mathcal{I}_k} \frac{\pi_k}{n_k} \mathrm{Pl}_t^{(i)} \mathrm{Pl}_t^* \\ &= \sum_{k\in\mathcal{Y}} L_{kl} \sum_{t=1}^{T} \pi_k \widehat{\mathrm{Pl}}_{kt} \mathrm{Pl}_t^*;\end{aligned}$$

where in these equations, we have

$$\widehat{\mathrm{Bel}}_{kt} = \widehat{\mathrm{Bel}}(\Phi(X) = \phi_t \mid Y = k) = \frac{1}{n_k} \sum_{i\in\mathcal{I}_k} \mathrm{Bel}_t^{(i)},$$

$$\widehat{\mathrm{Pl}}_{kt} = \widehat{\mathrm{Pl}}(\Phi(X) = \phi_t \mid Y = k) = \frac{1}{n_k} \sum_{i\in\mathcal{I}_k} \mathrm{Pl}_t^{(i)}.$$

These terms provide "average" belief and plausibility measures for each class k and profile ϕ_t. Intuitively, $\widehat{\mathrm{Bel}}_{kt}$ aggregates the masses assigned specifically to the profile ϕ_t, whereas $\widehat{\mathrm{Pl}}_{kt}$ accounts for all subsets of profiles that include ϕ_t. These measures define the lower and upper bounds on the available evidence for profile ϕ_t under class k, thereby quantifying the inherent uncertainty in the distribution.

By using a loss function L, we obtain a belief-based risk $\underline{\mathcal{R}_l}(x)$ and a plausibility-based risk $\overline{\mathcal{R}_l}(x)$, which jointly provide an interval-valued measure of the prediction risk. Although these values are not equivalent to the classical expected risk, they offer valuable insight into the uncertainty surrounding the classification decision. The former represents an optimistic estimate, reflecting the minimum possible risk under the assumption that only strongly supported evidence contributes to the decision. If $\underline{\mathcal{R}_l}(x)$ is high, it suggests that much of the available evidence definitively supports the choice of (ϕ_t, l), leading to a relatively higher lower bound. Conversely, $\overline{\mathcal{R}_l}(x)$ provides a pessimistic estimate, representing the highest possible risk when considering all evidence that might support (ϕ_t, l). If $\overline{\mathcal{R}_l}(x)$ is high, it implies that a substantial amount of evidence might support this association, leading to a larger worst-case risk estimate.

By leveraging these bounds and different decision criteria, we provide a structured approach to managing classification uncertainty, ensuring that final decisions are made with an appropriate balance between accuracy and determinacy in environments with high uncertainty.

4.3. Approximation of maximality. Using the approximation from the previous subsection, the optimization problem specified by maximality is as follows:

$$\begin{aligned}\min \sum_{k\in\mathcal{Y}} (L_{km} - L_{kn}) \sum_{t=1}^{T} \pi_k \widehat{p}_{kt} p_t^* > 0, \quad \forall l \in \mathcal{Y}, \\ \text{s.t.} \quad \mathrm{Bel}_t \le p_t \le \mathrm{Pl}_t.\end{aligned}$$

Here we use an approximation similar to the previous idea. It is not difficult to see that if $L_{km} - L_{kn} < 0$, the function to be optimized is negatively proportional to $\widehat{p}_{kt}\, p_t^*$, so its minimum (*i.e.*, the most adverse value for

Name	# Instances	# Feature	# classes
Iris	150	4	3
Breast Cancer	569	30	2
Wine	178	13	3
Glass	214	10	6
Customer	440	8	3
Seed	199	7	3

Table 1. *Information about the experimental datasets.*

the dominance condition) is obtained by setting the probabilities to their upper bounds:

$$\widehat{p}_{kt}\, p_t^* = \widehat{\mathrm{Pl}}_{kt}\, \mathrm{Pl}_t^*.$$

If $L_{km} - L_{kn} > 0$, the maximum is obtained by setting the probabilities to their lower bounds:

$$\widehat{p}_{kt}\, p_t^* = \widehat{\mathrm{Bel}}_{kt}\, \mathrm{Bel}_t^*$$

In fact, for the same class k, p_t^* needs to remain the same, but for computational efficiency, we can use this double approximation. This approximation reduces the computational complexity of full bilinear optimization to linear time, while still providing a meaningful and conservative evaluation of the maximality condition.

5. Experiments

To evaluate the performance of the proposed credal discrete classifier (CDC) and compare it with the soft-probabilistic discrete Bayesian classifier (SPDBC), we conduct experiments on six benchmark datasets from the UCI Machine Learning Repository as detailed in Table 1. The primary goals of our experiments are to compare CDC's interval-based risk estimates with the single-point estimates of SPDBC, and to analyze the trade-off between accuracy and cautiousness in set-valued predictions.

5.1. Setup. To ensure a fair comparison, both CDC and SPDBC are both built upon the same evidential C-Means (ECM) clustering framework. Since CDC requires belief mass functions, while SPDBC operates with probabilities, we derive the latter using the pignistic probability transformation [20]:

$$\mathrm{BetP}(\Phi(X) = \phi_t) = \sum_{\phi_t \in A} \frac{m(\Phi(X) = A)}{|A|}, \quad \forall \phi_t \in \Omega, \tag{13}$$

where $|A|$ represents the cardinality of focal set $A \subset \Omega$.

The Evidential C-Means algorithm [14] is an evidential version of fuzzy C-Means, controlled by three hyperparameters α, β and δ. Its objective function is as follows:

$$J_{\mathrm{ECM}} := \sum_{i=1}^{n} \sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} c_j^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \tag{14}$$

Here c_j is he cardinality factor associated with the focal set A_j, m_{ij} is the mass assignment of data point x_i to the set A_j and d_{ij} is the distance between data point x_i to the set A_j.The hyperparameter α is used to control the mass assigned to focal sets with high cardinality, β determines the fuzziness of the clustering, and δ controls the amount of mass given to the empty set. To analyze the influence of these parameters on CDC's performances, we particularly focus on variations in α, as it directly affects the distribution of mass across focal sets, thereby influencing the resulting risk intervals.

Obviously, as α increases, more mass is assigned to focal sets with lower cardinality, while focal sets with higher cardinality receive almost no mass. This leads to a reduction in the risk interval, and in extreme cases, the upper and lower bounds of the risk coincide, eliminating uncertainty and forcing the classifier to output a single class. Conversely, when α decreases, more mass is allocated to focal sets with higher cardinality, resulting in an excessively large risk interval and overly conservative decisions. Therefore, we aim to investigate how α affects the results.

In the full ECM algorithm, the mass function must be computed for $2^{|\Omega|}$ focal sets, leading to an exponential increase in the number of parameters to optimize with respect to the number of clusters (and a linear increase with the number of data points). However, to reduce computational complexity, one can restrict the size of focal set to a limited number. For instance, constraining focal sets to single classes or at most two-class combinations reduces complexity from $2^{|\Omega|}$ to $|\Omega|^2$, striking a balance between method flexibility and computational feasibility. In this experiment, we will limit the size of the focal set to no more than 2 (except Ω).

5.2. Metrics. Evaluating a credal or set-valued classifiers require metrics that balance between the size of the predictions, in terms of number of classes, and the accuracy of the predictions. A set-valued classifier should be rewarded if the observation is within the returned classes, and increasingly penalized as the number of classes increases, as it becomes less informative. A commonly used global evaluation metric to achieve this is discounted accuracy.

Let $U(x)$ denote the set of predicted classes for an instance x. Discounted accuracy assigns a reward of $1/|U(x)|$ if the true class y belongs to $U(x)$, thereby penalizing indeterminate predictions. In the binary classification case, discounted accuracy assigns a reward of $1/2$ to each indeterminate prediction (corresponding to a random selection between the two classes).

As emphasized by Zaffalon, Corani, and Mauá [25], when comparing precise classifiers and credal classifiers, using utility-discounted predictive accuracy measures can more fairly evaluate the performance of credible classifiers and highlight their value in high-stakes decisions.

The two most popular of these measures are arguably u_{65} and u_{80}:

$$u_{65}(z) = -0.6z^2 + 1.6z, \tag{15a}$$
$$u_{80}(z) = -1.2z^2 + 2.2z, \tag{15b}$$

with $z = 1/|U(x)|$. In binary classification, u_{65} rewards an indeterminate prediction with 0.65, and u_{80} with 0.80. In other words, u_{80} encourages indeterminate predictions more than u_{65}: this metric may therefore be more reasonable in some risk-averse scenarios such as medical diagnosis and autonomous driving. It is worth noting that for the precise classifier, since it can only output one class, its discounted accuracy will always be 1 or 0, so its u_{65} score and u_{80} score will be the same.

5.3. Procedure. To ensure statistical robustness, we perform 5×20 multiple cross-validation and follow these steps:

1. Divide the data set into a training set and a test set.
2. Use the ECM algorithm to cluster the training set to obtain the mass function $m(\Phi(x_i) = A)$.
3. Use the ECM parameters obtained in step 2 to cluster the test set to obtain the mass function $m(\Phi(x) = A)$.
4. Compute CDC's risk bounds and apply decision criteria to obtain classifications with 0-1 loss function.
5. Transform mass functions into pignistic probabilities and use SPDBC for classification.
6. Compare the performance of CDC and SPDBC using the evaluation metrics.

The experimental results are shown in Fig. 3 and Fig. 4. The difference between scores u_{65} and u_{80} under the same decision criteria can be seen as how many "sets" appear in the prediction. When the u_{65} and u_{80} scores are equal, it means that all predictions in the result are singletons. When the difference is large, it can be seen that there are many "set-valued" predictions.

5.4. Results. From the Figures 3 and 4, we can see that when α becomes larger, SPDBC (represented by BetP) and CDC are almost equivalent. This is because when α becomes larger, most of the mass is concentrated on singletons: in this case, we have $\mathrm{Bel}_t^{(i)} \approx \mathrm{Pl}_t^{(i)} \approx \mathrm{BetP}(\Phi(x_i) = \phi_t)$, and CDC degenerates into SPDBC.

When α becomes smaller, the situation will be reversed and the risk interval$[\underline{\mathcal{R}_l}(x), \overline{\mathcal{R}_l}(x)]$ will become larger. In this case, interval dominance is more difficult to achieve, so the output will gradually shift from precise decisions to completely imprecise ones. In extreme cases, interval dominance will output the full set Ω for each prediction point, while component-wise strict dominance is more robust to the change in the length of the risk interval. Therefore, for CDC, how to distribute the

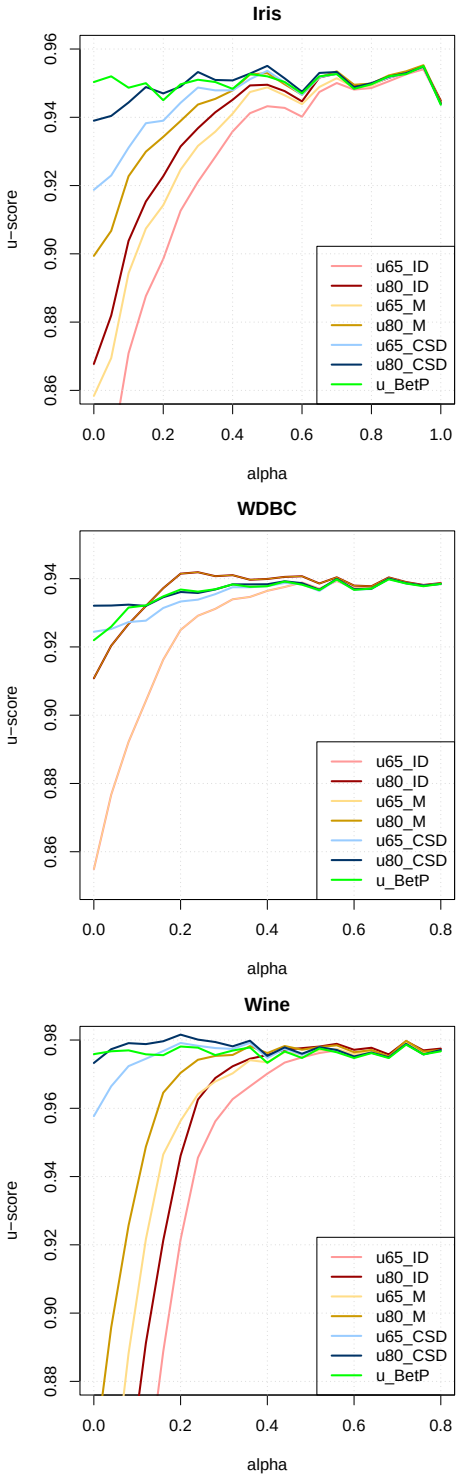

Figure 3. *U-scores for Iris, Breast Cancer and Wine with decision criteria interval dominance (ID), component-wise strict dominance (CSD), maximality (M) and pignistic probability (BetP).*

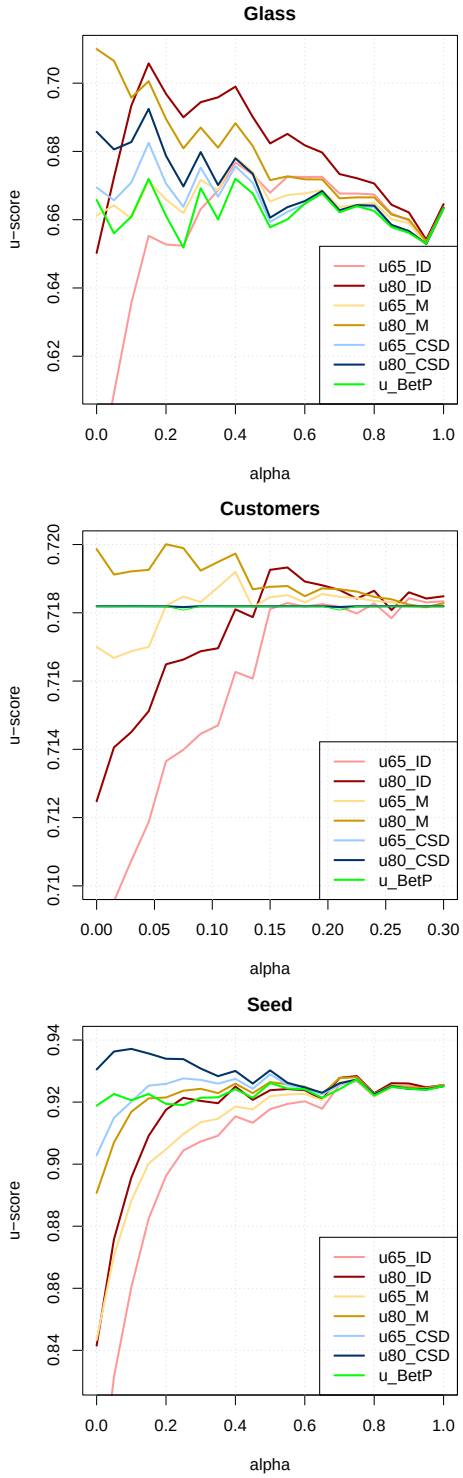

Figure 4. *U-scores for Glass, Customers and Seed with decision criteria interval dominance (ID), component-wise strict dominance (CSD), maximality (M) and pignistic probability (BetP).*

mass of focal sets with high cardinality is very important. Overall, the inclusion mentioned in Section 3.4 can be observed on the curves (for a given metric, ID curve is dominated by M curve, dominated in its turn by the CSD one).

To further interpret the experimental results, we observe that when the parameter α is appropriately chosen, all CDC variants based on component-wise strict dominance perform at least as well as SPDBC. Moreover, in all datasets except Breast Cancer, the u-score of SPDBC remains nearly constant, fluctuating around a fixed value. This stability arises from the fact that α only influences the mass assigned to focal sets with high cardinality, without altering the relative strengths of singleton profile assignments. Since SPDBC relies on the pignistic transformation, which uniformly distributes the mass of composite focal sets across the profiles they include, the resulting profile probabilities remain virtually unchanged as α varies.

For the Breast Cancer, Customers, and Seed datasets, we further observe that as α increases, the u-scores of CDC under various decision rules tend to increase at first, and then slightly decrease before stabilizing. This pattern reflects the transition from a highly cautious behavior—where a large portion of the mass is assigned to imprecise (high-cardinality) sets—to a nearly deterministic behavior where most mass is concentrated on singleton sets. The peak u-score is typically achieved at an intermediate α, indicating an optimal trade-off between accuracy and determinacy.

On the Glass dataset, CDC significantly outperforms SPDBC. This is largely due to the high number of classes in this dataset, which causes CDC to produce more set-valued predictions. Since the u-score rewards such predictions as long as the true class is included in the predicted set, CDC benefits more in this context.

For the Customers dataset, we note that the u-score varies very little across different α values, and that CDC effectively degenerates into SPDBC as early as $\alpha = 0.3$. This likely results from ECM perceiving a well-separated cluster structure in the data, with sharp boundaries between groups, which makes it difficult to assign significant mass to composite focal sets.

Overall, the results demonstrate that CDC provides a flexible trade-off between accuracy and precision, making it a valuable alternative to traditional probabilistic classifiers in uncertain environments.

6. Conclusions and Perspectives

We have introduced the credal discrete classifier (CDC), an extension of discrete Bayesian classifiers designed to remain robust under imperfect data. By employing belief functions to represent uncertain probabilities, CDC produces lower and upper bounds on the expected risk. This formulation naturally accommodates decision

criteria—such as interval dominance and componentwise strict dominance—that permit cautious, set-valued predictions when necessary.

Our empirical evaluations reveal that CDC performs on par with or better than its precise counterpart, the Soft Probabilistic Discrete Bayesian classifier (SPDBC). The risk-interval perspective also offers a more nuanced view of uncertainty, providing decision-makers with a controllable trade-off between precision (single-label outcomes) and caution (set-valued results).

Future research directions include refining the approximation schemes for risk-bound computation, comparing with other imprecise classifiers, and investigating the various solutions advocated in Section 3.2. Overall, this work underscores the value of embracing imprecision in probability estimates for enhanced robustness and interpretable decision-making in challenging classification scenarios.

Additional author information

Acknowledgements. The authors acknowledge the support of the French "Agence Nationale de la Recherche" (ANR), which funded this research under grant ANR-24-CE23-0847-01 (project DMC).

References

[1] Asmala Ahmad and Shaun Quegan. "Analysis of maximum likelihood classification on multispectral data". In: *Applied Mathematical Sciences* 6.129 (2012), pp. 6425–6436.

[2] James O. Berger. *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. New York: Springer, 1985.

[3] James C Bezdek, Robert Ehrlich, and William Full. "FCM: The fuzzy c-means clustering algorithm". In: *Computers & geosciences* 10.2-3 (1984), pp. 191–203.

[4] Wenlong Chen, Cyprien Gilet, Benjamin Quost, and Sébastien Destercke. "Robust Discrete Bayesian Classifier". In: *Scalable Uncertainty Management: 16th International Conference, SUM 2024, Palermo, Italy, November 27-29, 2024, Proceedings*. Vol. 15350. Springer Nature. 2024, p. 100.

[5] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22.

[6] T.S. Ferguson. *Mathematical Statistics : A Decision Theoretic Approach*. Academic Press, 1967.

[7] Cyprien Gilet. "Discrete minimax classifier for personalized diagnosis in medicine". PhD Thesis. Université Côte d'Azur, 2021. URL: https://tel.archives-ouvertes.fr/tel-03553934.

[8] Cyprien Gilet, Susana Barbosa, and Lionel Fillatre. "Discrete box-constrained minimax classifier for uncertain and imbalanced class proportions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.6 (2020), pp. 2923–2937.

[9] Daniel Grossman and Pedro Domingos. "Learning Bayesian network classifiers by maximizing conditional likelihood". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 46.

[10] George H John and Pat Langley. "Estimating continuous distributions in Bayesian classifiers". 2013. arXiv: 1302.4964.

[11] William R Klecka. *Discriminant analysis*. Sage, 1980.

[12] Scott Kolodziej, Pedro M Castro, and Ignacio E Grossmann. "Global optimization of bilinear programs with a multiparametric disaggregation technique". In: *Journal of Global Optimization* 57 (2013), pp. 1039–1063.

[13] Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press, 1980.

[14] Marie-Hélène Masson and Thierry Denoeux. "ECM: An evidential version of the fuzzy c-means algorithm". In: *Pattern Recognition* 41.4 (2008), pp. 1384–1397.

[15] Garth P McCormick. "Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems". In: *Mathematical programming* 10.1 (1976), pp. 147–175.

[16] Ignacio Montes, Enrique Miranda, and Sébastien Destercke. "Unifying neighbourhood and distortion models: part I–new results on old models". In: *International Journal of General Systems* 49.6 (2020), pp. 602–635.

[17] Ignacio Quesada and Ignacio E Grossmann. "A global optimization algorithm for linear fractional and bilinear programs". In: *Journal of Global Optimization* 6 (1995), pp. 39–76.

[18] Irina Rish et al. "An empirical study of the naive Bayes classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22. Seattle, WA, USA; 2001, pp. 41–46.

[19] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton university press, 1976.

[20] Philippe Smets. “The combination of evidence in the transferable belief model”. In: *IEEE Transactions on pattern analysis and machine intelligence* 12.5 (1990), pp. 447–458.

[21] Matthias CM Troffaes. “Decision making under uncertainty using imprecise probabilities”. In: *International journal of approximate reasoning* 45.1 (2007), pp. 17–29.

[22] Peter Walley. “Statistical reasoning with imprecise probabilities”. In: (1991).

[23] Danan Suryo Wicaksono and Iftekhar A Karimi. “Piecewise MILP under-and overestimators for global optimization of bilinear programs”. In: *AIChE Journal* 54.4 (2008), pp. 991–1008.

[24] Marco Zaffalon. “The naive credal classifier”. In: *Journal of statistical planning and inference* 105.1 (2002), pp. 5–21.

[25] Marco Zaffalon, Giorgio Corani, and Denis Mauá. “Evaluating credal classifiers by utility-discounted predictive accuracy”. In: *International Journal of Approximate Reasoning* 53.8 (2012), pp. 1282–1301.