# Imprecise Classification with Non-parametric Predictive Inference

Serafín Moral(✉), Carlos J. Mantas, Javier G. Castellano, and Joaquín Abellán

Department of Computer Science and Artifical Intelligence, University of Granada, Granada, Spain
{seramoral,cmantas,fjgc,jabellan}@decsai.ugr.es

**Abstract.** In many situations, classifiers predict a set of states of a class variable because there is no information enough to point only one state. In the data mining area, this task is known as Imprecise Classification. Decision Trees that use imprecise probabilities, also known as Credal Decision Trees (CDTs), have been adapted to this field. The adaptation proposed so far uses the Imprecise Dirichlet Model (IDM), a mathematical model of imprecise probabilities that assumes prior knowledge about the data, depending strongly on a hyperparameter. This strong dependence is solved with the Non-Parametric Predictive Inference Model (NPI-M), also based on imprecise probabilities. This model does not make any prior assumption of the data and does not have parameters. In this work, we propose a new adaptation of CDTs to Imprecise Classification based on the NPI-M. An experimental study carried out in this research shows that the adaptation with NPI-M has an equivalent performance than the one obtained with the adaptation based on the IDM with the best choice of the hyperparameter. Consequently, since the NPI-M is a non-parametric approach, it is concluded that the NPI-M is more appropriated than the IDM to be applied to the adaptation of CDTs to Imprecise Classification.

**Keywords:** Imprecise classification · Credal decision trees · IDM · NPI-M · Imprece probabilities

## 1 Introduction

Supervised classification [15] aims to predict the value of a *class variable* associated with an instance, described by a set of *features* or *attributes*. This prediction usually consists of a single value.

However, in many cases, there is no information available enough to point only one state of the class variable. In these cases, it is more informative that the classifier predicts a set of values of the class variable, which is known as an

Supported by the Spanish "Ministerio de Economía y Competitividad" and by "Fondo Europeo de Desarrollo Regional" (FEDER) under Project TEC2015-69496-R.

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1238, pp. 53–66, 2020.
https://doi.org/10.1007/978-3-030-50143-3_5

*imprecise prediction*. Classifiers that make this type of predictions are known as *imprecise classifiers*.

When it is used an imprecise classifier, a set of class values might be obtained. It is composed of those states for which there is no another "better" one according to a criterion, which is called *dominance criterion*. The set of predicted states of the class variable is known as the set of *non-dominated states*.

In order to build an imprecise classifier, it is more suitable to apply models based on imprecise probabilities, instead of the ones that use the classical probability theory. In the literature, there are many mathematical theories associated with imprecise probabilities, such as belief functions, closed and convex sets of probability distributions (also called credal sets), probability intervals, etc [16].

In the literature, few methods for imprecise classification have been developed. The first one of them was the Naive Credal Classifier (NCC) [10,24]. It uses the Imprecise Dirichlet Model (IDM) [22], a mathematical model of imprecise probabilities that makes statistical inferences from multinomial data, and the Naive Bayes assumption (all the attributes are independent given the class variable) to produce an imprecise classification.

In [4], it is proposed a new adaptation of the Credal Decision Trees (CDTs) [5], very simple and interpretable models, to Imprecise Classification. It is called Imprecise Credal Decision Tree (ICDT). In that work, it is shown, via an experimental analysis, that ICDT is a more informative method than NCC since it is more precise. In this work, we focus on the ICDT algorithm.

The ICDT proposed so far is based on the IDM. This model satisfies several principles which have been claimed to be desirable for inference, such as the *representation invariance principle* [22]. According to it, inferences on future events should be independent of the arrangement and labeling of the sample space. Nevertheless, IDM assumes previous knowledge about the data through a single hyperparameter $s$ [22]. It is not a very desirable property because these assumptions are not always realistic.

For the previous reason, a Non Parametric model for Predictive Inference (NPI-M) was proposed in [8]. This model does not make any prior assumptions about the data. In addition, NPI-M is a nonparametric approach.

Both IDM and NPI-M have been applied to Decision Trees (DT) for precise classification in the literature [6,18,19]. When the IDM is applied to DTs, it has been shown that the performance has a strong dependence on the $s$ parameter [19]. In [6], the NPI-M is shown to have always an equivalent performance to IDM with the standard s value when both models are applied to DTs.

For the previous reasons, in this work, we propose a new adaptation of CDTs to Imprecise Classification based on the NPI-M. It is called Imprecise Credal Decision Tree NPI (ICDT-NPI). It is similar to the already existing adaptation, but our proposed one is based on the NPI-M, instead of the IDM.

An extensive experimental research is carried out in this work. In it, we use the ICDT-NPI algorithm and the ICDT with different values of the $s$ parameter for the IDM. This experimentation shows that, as in precise classification, the NPI-M provides equivalent results to the IDM with the best choice of the $s$

hyperparameter when both models are applied to the adaptation of CDTs to Imprecise Classification.

This paper is arranged as follows: The Imprecise Dirichlet Model and the Non-Parametric Predictive Inference Model are explained in Sects. 2 and 3, respectively. Section 4 describes the dominance criteria for Imprecise Classification used in this research. The adaptation of the Credal Decision Trees to Imprecise Classification is exposed in Sect. 5. Section 6 describes the main evaluation metrics that are used in imprecise classification. In Sect. 7, the experimental analysis is detailed. Conclusions are given in Sect. 8.

## 2 The Imprecise Dirichlet Model

Let us suppose that we have a dataset $\mathcal{D}$ with $N$ instances. Let $X$ be an attribute that takes values in $\{x_1, \cdots, x_t\}$.

The Imprecise Dirichlet Model (IDM) [22] is subsumed into the probability intervals theory [11]. According to this model, the variable $X$ takes each one of its possible values $x_i$, $1 \leq i \leq t$ with a probability that belongs to the following interval:

$$I_i = \left\{ \left[ \frac{n_i}{N+s}, \frac{n_i+s}{N+s} \right] \right\}, \forall i = 1, 2, \ldots, t, \tag{1}$$

being $n_i$ the number of instances in $\mathcal{D}$ for which $X = x_i$, $\forall i = 1, 2, \ldots, t$ and $s > 0$ a given parameter of the model.

As it is shown in [1], this set of probability intervals is reachable and gives rise to the following closed and convex set of probability distributions, also called credal set:

$$\mathcal{P}^{\mathcal{D}}(X) = \left\{ p \mid \sum_{i=1}^{t} p(x_i) = 1,\, p(x_i) \in I_i, \quad \forall i = 1, 2, \ldots, t \right\}. \tag{2}$$

A crucial issue is the selection of the $s$ hyperparameter. It is easy to observe that, if the $s$ value is higher, then the intervals are wider. This parameter determines the speed of convergence of lower and upper probabilities as the size of the training set is larger. In [22], the values $s = 1$ and $s = 2$ are proposed.

## 3 Non-parametric Predictive Inference Model

Let $X$ be a discrete variable whose set of possible values is $\{x_1, \cdots, x_T\}$. Let us suppose that there is a sample of N independent and identically distributed outcomes of $X$. Let $n_i$ be the number of observations for which $X = x_i$, $\forall i = 1, 2, \ldots, T$. Let us assume that the first $t$ observations have been observed, where $1 \leq t \leq T$, which implies that $n_i > 0$, $\forall i = 1, 2, \cdots, t$ and $n_i = 0$, $\forall i = t+1, \cdots, T$. Clearly, $\sum_{i=1}^{t} n_i = N$.

The Non-Parametric Predictive Inference Model (NPI-M) [8,9] utilizes a *probability wheel* representation of the data. On it, it is used a line from the center of the wheel to its boundary to represent each one of the observations. The wheel is partitioned into $N$ slices with the same size. Each possible value can be represented only by a single sector of the wheel. This implies that two or more lines representing the same category must always be positioned next to each other on the wheel. The NPI-M is based on the circular $A_{(n)}$ assumption [9]. According to it, the probability that the next observation falls into any given slice is $\frac{1}{N}$. Thus, it must be decided which value of the $X$ variable represents. If two lines that represent the same category border to a slice, that slice must be assigned to this value. Nevertheless, when a slice is bordered by two lines that represent different values, it can be assigned to one of the two categories associated with the slice's bordering lines, or to any value that has not been observed yet.

Let $A \subseteq \{x_1, x_2, \dots, x_T\}$ be a subset of the set of possible values of the $X$ variable. Let us denote $n_A = \sum_{x_i \in A} n_i$ the number of outcomes of $X$ for which its value belongs to $A$ and $r_A = |\{x_i \in A \mid n_i > 0,\ 1 \leq i \leq t\}|$ the number of possible values in $A$ that have been already observed.

In order to determine the lower and upper probabilities of $A$, NPI-M considers all the possible configurations of the wheel. The difference between both probabilities is due to the non-observed categories. In [3], it is shown that the lower and upper probabilities of $A$ are obtained as follows:

$$P_*(A) = \frac{n_A - \min(r_A, |\overline{A}|)}{N}, \quad P^*(A) = \frac{n_A + \min(|A|, t - r_A)}{N}. \tag{3}$$

As it can be seen, for singletons, $\{x_i\}$, $1 \leq i \leq T$, the lower and upper probabilities are given by:

$$P_*(\{x_i\}) = \max\left(\frac{n_i - 1}{N}, 0\right), \quad P^*(\{x_i\}) = \min\left(\frac{n_i + 1}{N}, 1\right).$$

Hence, it is disposed of the following set of probability intervals for singletons:

$$\mathcal{I} = \left\{[l_i, u_i],\ l_i = \max\left(\frac{n_i - 1}{N}, 0\right),\ u_i = \min\left(\frac{n_i + 1}{N}, 1\right), \quad \forall i = 1, \dots, T\right\}.$$

According to [11], this set of probability intervals corresponds to the following credal set:

$$\mathcal{P}(\mathcal{I}) = \{p \in \mathcal{P}(X) \mid p(x_i) \in [l_i, u_i], \quad \forall i = 1, 2 \dots, T\}. \tag{4}$$

being $\mathcal{P}(X)$ the set of all probability distributions on the $X$ variable, $l_i = \max\left(\frac{n_i - 1}{N}, 0\right)$, and $u_i = \min\left(\frac{n_i + 1}{N}, 1\right)$, $\forall i = 1, 2, \dots, T$.

In [3], it was proved that the lower and upper probabilities associated with $\mathcal{P}(\mathcal{I})$ coincide with the lower and upper probabilities given by (3). Therefore, the lower and upper probabilities corresponding to the NPI-M can be extracted via the lower and upper probabilities for singletons, which produce a set of probability intervals, and, consequently, a credal set. Nevertheless, in [3], it is shown that, in this set, there are probability distributions that are not compatible with the NPI-M.

If we consider all the probability distributions belonging to $\mathcal{P}(\mathcal{I})$, it is obtained an approximated model, called Approximate Non-Parametric Predictive Inference Model (A-NPI-M) [3]. It utilizes the convex hull of the set of probability distributions compatible with the NPI-M. In this way, when the A-NPI-M is used, a set of difficult constraints is avoided and the exact model is simplified. In [3], it is shown that NPI-M and A-NPI-M have a similar behavior when both models are applied to CDTs. For these reasons, in this work, we consider the A-NPI-M.

## 4 Dominance Criteria in Imprecise Classification

In Imprecise Classification, it is used a *dominance criterion* to select the states of the class variable that are not "defeated" under that criterion by another. In order to do it, if we have a set of probability intervals, as in this research, we can use the bounds of the intervals.

Let $c_i$ and $c_j$ be two possible values of the class variable $C$. Two *dominance criteria* very used are the following:

1. Let $[l_i, u_i]$ and $[l_j, u_j]$ be the probability intervals on how $c_i$ and $c_j$ happen, respectively. It is said that there is *stochastic dominance* or *strong dominance* of $c_i$ on $c_j$ if, and only if, $l_i \geq u_j$.
2. Let suppose now that the probability of the class variable $C$ is expressed by a non-empty credal set $\mathcal{P}$. It is said that there is credal dominance of $c_i$ on $c_j$ iff $p(C = c_i) \geq p(C = c_j)$, for all probability distribution $p \in \mathcal{P}$.

Credal dominance is a more significant criterion than stochastic dominance [24]. However, it is usually more difficult to verify. Under the IDM and the A-NPI-M, both dominance criteria are equivalent [2]. Hence, with both IDM and A-NPI-M, if we check that one state dominates stochastically to another, then we know that there is credal dominance of the first state on the second one. Therefore, with IDM, as well as with A-NPI-M, it is just necessary to consider the extreme values of the intervals to know the cases of credal dominance among the possible states of $C$.

## 5 Credal Decision Trees for Imprecise Classification

The adaptation of Credal Decision Tree algorithm (CDT) [5] to Imprecise Classification (ICDT) was proposed in [4]. It is called Imprecise Credal Decision Tree algorithm (ICDT).

As in CDTs, in ICDT, each node corresponds to an attribute or feature variable and there is a branch for each possible value of that attribute. When entering a feature in a node does not provide more information about the class variable according to a criterion, a terminal or leaf node is reached. Unlike in precise classification, where this leaf node is labeled with the most probable class value, in ICDTs, the leafs nodes are empty. When a new example is required to be classified, it is made a path from the root to a terminal node using the values of its attributes. Whereas in precise classification the most probable class value in that leaf node is assigned, in ICDTs, for each value of the class variable, it is obtained a probability interval.

The most important issue of the building process of the ICDT is the split criterion, i.e, the criterion utilized to select the attribute to split in each node. In ICDT, the split criterion is the same as the one used in CDT.

Let us suppose that $\mathcal{D}$ is a partition of the training set in a certain node. Let $C$ be the class variable and let us suppose that $\{c_1, \ldots, c_k\}$ are its possible values. Let $X$ be an attribute variable and let $\{x_1, x_2, \ldots, x_t\}$ be its possible values. Let us assume that $\mathcal{P}^{\mathcal{D}}(C)$ is the credal set on $\mathcal{D}$ associated with $C$ corresponding to a model based on probability intervals[1] .

The split criterion utilized in the ICDT algorithm utilizes the maximum of the Shannon entropy [21] on $\mathcal{P}^{\mathcal{D}}(C)$:

$$H^*(\mathcal{P}^{\mathcal{D}}(C)) = \max\left\{H(p) \mid p \in \mathcal{P}^{\mathcal{D}}(C)\right\} \tag{5}$$

being $H$ the Shannon entropy.

The maximum of entropy is a well-established measure on credal sets that satisfies good properties and behavior [16].

Hence, the split criterion used in ICDT is the Imprecise Information Gain (IIG) [5]. It is defined as follows:

$$IIG(C, X) = H^*(\mathcal{P}^{\mathcal{D}}(C)) - \sum_{i=1}^{t} P^{\mathcal{D}}(X = x_i) H^*(\mathcal{P}^{\mathcal{D}}(C \mid X = x_i)), \tag{6}$$

where $P^{\mathcal{D}}(X)$ is the maximum of entropy on the credal set corresponding to the $X$ attribute and $H^*(\mathcal{P}^{\mathcal{D}}(C \mid X = x_i))$ is the maximum of the entropy on the credal set associated with the $C$ variable and with the partition of $\mathcal{D}$ composed by the instances of $\mathcal{D}$ that verify that $X = x_i$.

The main difference among the CDT and ICDT algorithms resides in the criterion utilized to classify an instance once a terminal node is reached. The CDT algorithm assigns the most frequent class value in that leaf. Nevertheless, the ICDT algorithm assigns a probability interval to each one of the possible values of the class variable using the relative frequencies in the leaf node and a model based on probability intervals. Then, a dominance criterion is used to obtain the set of non-dominated states. In this research, the models considered are the IDM and the A-NPI-M. Thus, since with these models stochastic and credal dominance are equivalent and the first one is much easier to verify, we use the stochastic dominance in this work.

The procedure to classify a new instance in the ICDT algorithm can be summarized in Fig. 1.

[1] In this work we will consider the A-NPI-M, unlike in [4], where the IDM is employed.

```
Procedure **Classify_Instance_ICDT**(Built Tree $\mathcal{T}$, new instance $\mathbf{x}$)

1. Apply $\mathbf{x}$ in $\mathcal{T}$ to reach a leaf node.
2. Obtain the probability intervals in this terminal node for $\mathbf{x}$, based on relative frequencies using a model based on imprecise probabilities:

   $\{[l_i, u_i], i = 1, \cdots, k\}$.

3. Apply a dominance criterion to the above intervals to get a set of non-dominated states for $\mathbf{x}$: $\{c_{i_1}, c_{i_2}, \cdots, c_{i_r}\}$, with $r \leq k$.
```

**Fig. 1.** Classification of a new instance in ICDT algorithm.

In this research, we compare the use of the A-NPI-M for the credal sets associated with the class variable, in the building process of the ICDT algorithm and to obtain the probability intervals for the class values in the terminal nodes, with the use of the IDM with different values of the $s$ parameter.

## 6 Evaluation Metrics in Imprecise Classification

An evaluation measure for Imprecise Classification should take into consideration two points. The first one of them is if the prediction is right, i.e if the real class value is among the predicted ones. The second point is how informative is the predicted set of states, which is measured by its cardinality.

Several metrics only focus on one of the issues commented above, such as:

- **Determinacy:** It is the proportion of instances for which the classifier returns a single class value.
- **Single Accuracy:** It consists of the accuracy between the instances for which there is just one predicted state.
- **Set Accuracy:** It measures, on the instances for which there is more than one predicted state, the proportion of them for which its real state is among the predicted ones.
- **Indeterminacy Size:** It is the average size of the predicted states set.

As it can be observed, none of these metrics is suitable to measure the whole performance of an imprecise classifier.

In [10], it was proposed a measure to provide a global evaluation of an imprecise classifier, called *Discounted Accuracy measure* (DACC), defined as:

$$DACC = \frac{1}{N_{Test}} \sum_{i=1}^{N_{Test}} \frac{(correct)_i}{|U_i|}, \tag{7}$$

where $N_{Test}$ is the number of instances of the test set; $U_i$ is the predicted states set for the i-th instance; $|U_i|$ its cardinality; $(correct)_i$ is equal to 1 if $U_i$ contains the real class value and 0 otherwise, $\forall i = 1, 2, \ldots, n$.

We shall denote $k$ the number of class values.

It can be observed that DACC is an accuracy measure: it does not add any value for the erroneous predictions and, for the correct ones, the added value is "penalized" by the number of predicted states. The optimal value of $DACC$ is 1. It is achieved when there is always a single predicted state and all the predictions are right. If the classifier always predicts all the possible values of the class variable, the value of DACC is $\frac{1}{k}$. This value should be lower because in this case, the classifier is not informative.

In [4], a new metric for imprecise classification, MIC, was proposed. It penalizes the errors in a strict sense. When there is a correct prediction for an instance, MIC adds a value that depends on $\frac{|U_i|}{k}$. If the prediction for an instance is incorrect, MIC adds a constant value, which depends on $k$. More specifically, MIC is defined as follows:

$$MIC = \frac{1}{N_{Test}} \left( \sum_{i:Success} \log \frac{|U_i|}{k} + \frac{1}{k-1} \sum_{i:Error} \log k \right) \tag{8}$$

As can be seen, the optimal value of MIC is $\log k$. It is reached when, for all the instances, only the real class value is predicted. Besides, when a classifier always returns as predicted states all the possible ones, i.e, when $|U_i| = k$, $\forall i = 1, \ldots, N_{Test}$, the value of $MIC$ is equal to 0. It makes sense because, in this case, the classifier does not give any information.

# 7 Experimentation

## 7.1 Experimental Settings

Remark that, in this experimentation, the aim is to compare the use of the A-NPI-M versus the IDM in the ICDT algorithm, in the building process and for the selection of non-dominated states. For the reasons explained in Sect. 5, the stochastic dominance criterion is used. For evaluation, we use the DACC and MIC measures, as in [4].

Within this Section, we call ICDT-IDM to the Imprecise Decision Tree with the IDM and ICDT-NPI to the Imprecise Decision Tree with the A-NPI-M.

To compare the performance between both algorithms, as in the experimentation carried out in [4], 34 known datasets have been used. They can be downloaded from the *UCI Machine Learning repository* [17]. These datasets are diverse concerning the size of the set, the number of continuous and discrete attributes, the number of values per variable, the number of class values, etc. The most relevant characteristics of each dataset can be seen in Table 1.

As it was done in [4], in each dataset, missing values have been replaced with the mean value for continuous variables and with modal values for discrete attributes. After that, in each database, continuous attributes have been discretized using the Fayyad and Irani's discretization method [13].

**Table 1.** Data set description. Column "N" is the number of instances in the data sets, column "Feat" is the number of features or attribute variables, column "Num" is the number of numerical variables, column "Nom" is the number of nominal variables, column "k" is the number of cases or states of the class variable (always a nominal variable) and column "Range" is the range of states of the nominal variables of each data set.

| Data set | N | Feat | Num | Nom | $k$ | Range |
|---|---|---|---|---|---|---|
| anneal | 898 | 38 | 6 | 32 | 6 | 2–10 |
| arrhythmia | 452 | 279 | 206 | 73 | 16 | 2 |
| audiology | 226 | 69 | 0 | 69 | 24 | 2–6 |
| autos | 205 | 25 | 15 | 10 | 7 | 2–22 |
| balance-scale | 625 | 4 | 4 | 0 | 3 | – |
| car | 1728 | 6 | 0 | 6 | 4 | 3–4 |
| cmc | 1473 | 9 | 2 | 7 | 3 | 2–4 |
| dermatology | 366 | 34 | 1 | 33 | 6 | 2–4 |
| ecoli | 366 | 7 | 7 | 0 | 7 | – |
| flags | 194 | 30 | 2 | 28 | 8 | 2–13 |
| hypothyroid | 3772 | 30 | 7 | 23 | 4 | 2–4 |
| iris | 150 | 4 | 4 | 0 | 3 | – |
| letter | 20000 | 16 | 16 | 0 | 26 | – |
| lymphography | 146 | 18 | 3 | 15 | 4 | 2–8 |
| mfeat-pixel | 2000 | 240 | 0 | 240 | 10 | 4–6 |
| nursery | 12960 | 8 | 0 | 8 | 4 | 2–4 |
| optdigits | 5620 | 64 | 64 | 0 | 10 | – |
| page-blocks | 5473 | 10 | 10 | 0 | 5 | – |
| pendigits | 10992 | 16 | 16 | 0 | 10 | – |
| postop-patient-data | 90 | 9 | 0 | 9 | 3 | 2–4 |
| primary-tumor | 339 | 17 | 0 | 17 | 21 | 2–3 |
| segment | 2310 | 19 | 16 | 0 | 7 | – |
| soybean | 683 | 35 | 0 | 35 | 19 | 2–7 |
| spectrometer | 531 | 101 | 100 | 1 | 48 | 4 |
| splice | 3190 | 60 | 0 | 60 | 3 | 4–6 |
| sponge | 76 | 44 | 0 | 44 | 3 | 2–9 |
| tae | 151 | 5 | 3 | 2 | 3 | 2 |
| vehicle | 946 | 18 | 18 | 0 | 4 | – |
| vowel | 990 | 11 | 10 | 1 | 11 | 2 |
| waveform | 5000 | 40 | 40 | 0 | 3 | – |
| wine | 178 | 13 | 13 | 0 | 3 | – |
| zoo | 101 | 16 | 1 | 16 | 7 | 2 |

We have used the Weka software [23] for this experimentation. We have started from the implementation of the ICDT-IDM algorithm given in this software and we have added the necessary methods to use the ICDT-NPI. For ICDT-IDM, three values of the s parameter have been used: $s = 1$, $s = 2$ and $s = 3$. The rest of the parameters used in both algorithms have been the ones given by default in Weka. This software has been also used for the preprocessing steps described above. We denote ICDT-IDMi to ICDT-IDM with $s = i$, for $i = 1, 2, 3$.

For each dataset, a 10-fold cross-validation procedure has been repeated 10 times.

For statistical comparisons, consistently with [12], we have used the following tests to compare more than two classifiers on a large number of datasets with a level of significance of $\alpha = 0.05$:

- **Friedman test** [14]: A non-parametric test that ranks the algorithms separately for each dataset (the best performing algorithm is assigned to the rank 1, the second-best, rank 2, and so on). The null hypothesis of the Friedman test is that all the algorithms have equivalent performance.
- When the null hypothesis of the Friedman test is rejected, all the algorithms are compared to each other by using the **Nemenyi test** [20].

For the statistical tests, the Keel software [7] has been used.

### 7.2 Results and Discussion

Tables 2 and 3 show, respectively, the main results corresponding to DACC and MIC measures. Specifically, these tables allow us to see the average values, the Friedman ranks, and the pairs of algorithms for which there are significant differences according to Nemenyi pos-hoc. We do not show the complete results here due to the limitations of space, they can be found in http://flanagan.ugr.es/IPMU2020.html.

**Table 2.** Summary of the results for the DACC measure. Column "Nemenyi" shows the algorithms in which the algorithm in the row performs significantly better according to the Nemenyi test.

| Algorithm | Average | Friedman rank | Nemenyi |
|---|---|---|---|
| ICDT-NPI | 0.7675 | 1.9118 | ICDT-IDM3 |
| ICDT-IDM1 | **0.7763** | 2.3382 | ICDT-IDM3 |
| ICDT-IDM2 | 0.7606 | 2.4853 | – |
| ICDT-IDM3 | 0.7482 | 3.2647 | – |

As it can be observed, for both DACC and MIC metrics, the best average value is obtained by ICDT-IDM1, followed by ICDT-NPI, ICDT-IDM2, and

**Table 3.** Summary of the results for the MIC measure. Column "Nemenyi" shows the algorithms in which the algorithm in the row performs significantly better according to the Nemenyi test.

| Algorithm | Average | Friedman rank | Nemenyi |
|---|---|---|---|
| ICDT-NPI | 1.3414 | 1.9706 | ICDT-IDM3 |
| ICDT-IDM1 | **1.3652** | 2.4412 | – |
| ICDT-IDM2 | 1.3334 | 2.5 | – |
| ICDT-IDM3 | 1.3065 | 3.0882 | – |

ICDT-IDM3. In addition, for both evaluation measures, the ICDT-NPI algorithm obtains the best rank according to the Friedman test. Regarding ICDT-IDM, the higher is the value of the $s$ parameter, the higher is the rank. The results of the Nemenyi post-hoc allow us to observe that the results obtained by ICDT-NPI are significantly better than the ones obtained by ICDT-IDM with the worst $s$ value ($s = 3$) for both MIC and DACC metrics. Also, for DACC, ICDT-IDM with $s = 1$ performs significantly better than ICDT-IDM with $s = 3$. For both evaluation metrics, ICDT-NPI, ICDT-IDM1, and ICDT-IDM2 have an equivalent performance.

Hence, the performance of the ICDT-IDM algorithm depends on the choice of the $s$ hyperparameter. Regarding ICDT-NPI, the results obtained for this algorithm are statistically equivalent to the ones obtained by ICDT-IDM with the best $s$ parameter. Furthermore, ICDT-NPI performs significantly better than ICDT-IDM with the worst value of the $s$ hyperparameter.

For a deeper analysis, Table 4 shows the average values of Determinacy, Single Accuracy, Set Accuracy and Indeterminacy size for each algorithm.

**Table 4.** Average results obtained for basic metrics by each algorithm. Best scores are marked in bold.

| Algorithm | Determinacy | Single accuracy | Set accuracy | Indeterminacy size |
|---|---|---|---|---|
| ICDT-NPI | 0.9002 | **0.8237** | **0.9561** | 7.9381 |
| ICDT-IDM1 | **0.9477** | 0.8023 | 0.8844 | **5.2955** |
| ICDT-IDM2 | 0.8985 | 0.8119 | 0.9168 | 5.9313 |
| ICDT-IDM3 | 0.8666 | 0.8151 | 0.9218 | 6.1346 |

Firstly, ICDT-IDM1 achieves the highest average Determinacy. It means that the highest number of instances for which only one state is predicted is obtained with ICDT-IDM1. NPI-M obtains the second-highest value in Determinacy, followed by ICDT-IDM2 and ICDT-IDM3 for all the noise levels.

However, for the accuracy among the instances for which it is predicted just a state of the class variable (Single Accuracy), ICDT-IDM obtains the worst performance with $s = 1$. In the ICDT-IDM algorithm, the higher is the value of the $s$ parameter, the better is the performance in Single Accuracy. The best Single Accuracy is obtained with ICDT-NPI.

Regarding Set Accuracy, which measures the average number of instances for which the real class value is between the predicted ones, the results are similar to the ones obtained in Single Accuracy: ICDT-IDM performs better as the value of the $s$ parameter is higher and ICDT-NPI outperforms ICDT-IDM with the three $s$ values considered.

The lowest value of the Indeterminacy size, which measures the average size of the non-dominated states, is achieved with ICDT-IDM1. Moreover, the lower is the $s$ value for ICDT-IDM, the lower is the indeterminacy size. The highest average number of non-dominated states is obtained with ICDT-NPI.

Therefore, with ICDT-NPI, it is attained the best trade-off between predicting only one state and making correct predictions. This algorithm obtains the second-highest score in Determinacy and the best one in Single Accuracy, whereas ICDT-IDM1, which achieves the highest Determinacy, obtains the worst results in Single Accuracy. Besides, when there is more than one predicted state, in the ICDT algorithm, the size of the predicted states sets is larger as the value if the $s$ value is higher and the largest set is obtained with the ICDT-NPI algorithm. Nevertheless, ICDT-NPI obtains the highest percentage of instances for which the real class value is predicted and, in the ICDT-IDM algorithm, this percentage is lower as the value of the $s$ parameter is higher.

**Summary of the Results:** The ICDT-IDM algorithm predicts the real class value more frequently as long as the value of the $s$ parameter is higher. However, if the $s$ value is higher, then the predictions made by ICDT-IDM are less informative in the sense that the size of the predicted class values set is larger. With ICDT-NPI, although the size of the predicted states set is, on average, larger than with ICDT-IDM, it is achieved the best trade-off between predicting fewer states of the class variable and making correct predictions.

The results obtained with DACC and MIC measures allow us to deduce that ICDT-NPI performs equivalently to ICDT-IDM with the best choice of the $s$ parameter. Moreover, the results obtained by ICDT-NPI are significantly better than the ones obtained by ICDT-IDM with the worst $s$ value. Consequently, the NPI-M is more suitable to be applied to the adaptation of CDTs to Imprecise Classification, since the NPI-M is free of parameters.

## 8 Conclusions

In this work, we have dealt with the problem of Imprecise Classification. Specifically, we have considered the adaptation of the Credal Decision Trees, Decision Trees that use imprecise probabilities, to this field.

The adaptation of Credal Decision Trees to Imprecise Classification proposed so far was based on the Imprecise Dirichlet Model, a model based on imprecise probabilities that assumes prior knowledge about the data through a hyperparameter $s$. For this reason, in this research, a new adaptation of Credal Decision Trees to Imprecise Classification based on the Non-Parametric Predictive Inference Model has been presented. This model, which is also based on imprecise probabilities, solves the main drawback of the Imprecise Dirichlet Model: it does not make any prior assumption about the data; it is a non-parametric approach.

An experimental research carried out in this work has shown that the new adaptation of Credal Decision Trees to Imprecise Classification based on the Non-Parametric Predictive Inference Model has equivalent performance to the Imprecise Credal Decision Tree based on the Imprecise Dirichlet Model with the best $s$ value. The results obtained by Imprecise Credal Decision Tree with Non-Parametric Predictive Inference Model are also significantly better than the ones obtained by Imprecise Credal Decision Tree with the Imprecise Dirichlet Model with the worst $s$ value. Although with the Non-Parametric Predictive Inference Model the set of predicted class values is larger than with the Imprecise Dirichlet Model, with the first model it is achieved a better trade-off between making correct predictions and predicting fewer states of the class variable.

Therefore, taking into account that the Non-Parametric Predictive Inference Model is free of parameters, it can be concluded that this model is more suitable to be applied to the adaptations of Credal Decision Trees to Imprecise Classification than the Imprecise Dirichlet Model.

## References

1. Abellán, J.: Uncertainty measures on probability intervals from the imprecise dirichlet model. Int. J. Gen. Syst. **35**(5), 509–528 (2006). https://doi.org/10.1080/03081070600687643
2. Abellán, J.: Equivalence relations among dominance concepts on probability intervals and general credal sets. Int. J. Gen. Syst. **41**(2), 109–122 (2012). https://doi.org/10.1080/03081079.2011.607449
3. Abellán, J., Baker, R.M., Coolen, F.P.: Maximising entropy on the nonparametric predictive inference model for multinomial data. Eur. J. Oper. Res. **212**(1), 112–122 (2011). https://doi.org/10.1016/j.ejor.2011.01.020
4. Abellán, J., Masegosa, A.R.: Imprecise classification with credal decision trees. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **20**(05), 763–787 (2012). https://doi.org/10.1142/S0218488512500353
5. Abellán, J., Moral, S.: Building classification trees using the total uncertainty criterion. Int. J. Intell. Syst. **18**(12), 1215–1225 (2003). https://doi.org/10.1002/int.10143
6. Abellán, J., Baker, R.M., Coolen, F.P., Crossman, R.J., Masegosa, A.R.: Classification with decision trees from a nonparametric predictive inference perspective. Comput. Stat. Data Anal. **71**, 789–802 (2014). https://doi.org/10.1016/j.csda.2013.02.009
7. Alcalá-Fdez, J., et al.: KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Comput. **13**(3), 307–318 (2009). https://doi.org/10.1007/s00500-008-0323-y

8. Coolen, F.P.A.: Learning from multinomial data: a nonparametric predictive alternative to the imprecise dirichlet model. In: Cozman, F.G., Nau, R., Seidenfeld, T. (eds.) ISIPTA 2005: Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications, pp. 125–134 (2005). (Published by SIPTA)
9. Coolen, F., Augustin, T.: A nonparametric predictive alternative to the imprecise Dirichlet model: The case of a known number of categories. Int. J. Approximate Reasoning **50**(2), 217–230 (2009). https://doi.org/10.1016/j.ijar.2008.03.011. Special Section on The Imprecise Dirichlet Model and Special Section on Bayesian Robustness (Issues in Imprecise Probability)
10. Corani, G., Zaffalon, M.: Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2. J. Mach. Learn. Res. **9**, 581–621 (2008)
11. De Campos, L.M., Huete, J.F., Moral, S.: Probability intervals: a tool for uncertain reasoning. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **02**(02), 167–196 (1994). https://doi.org/10.1142/S0218488594000146
12. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
13. Fayyad, U., Irani, K.: Multi-valued interval discretization of continuous-valued attributes for classification learning. In: Proceeding of the 13th International Joint Conference on Artificial Inteligence, pp. 1022–1027. Morgan Kaufmann (1993)
14. Friedman, M.: A comparison of alternative tests of significance for the problem of $m$ rankings. Ann. Math. Stat. **11**(1), 86–92 (1940). https://doi.org/10.1214/aoms/1177731944
15. Hand, D.J.: Construction and Assessment of Classification Rules. Wiley, New York (1997)
16. Klir, G.J.: Uncertainty and Information: Foundations of Generalized Information Theory. Wiley (2006). https://doi.org/10.1002/0471755575
17. Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml
18. Mantas, C.J., Abellán, J.: Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. Expert Syst. Appl. **41**(10), 4625–4637 (2014). https://doi.org/10.1016/j.eswa.2014.01.017
19. Mantas, C.J., Abellán, J., Castellano, J.G.: Analysis of Credal-C4.5 for classification in noisy domains. Expert Syst. Appl. **61**, 314–326 (2016). https://doi.org/10.1016/j.eswa.2016.05.035
20. Nemenyi, P.: Distribution-free multiple comparisons. Doctoral dissertation, Princeton University, New Jersey, USA (1963)
21. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
22. Walley, P.: Inferences from multinomial data; learning about a bag of marbles (with discussion). J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 3–57 (1996). https://doi.org/10.2307/2346164
23. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems, 2nd edn. Morgan Kaufmann Publishers Inc., San Francisco (2005)
24. Zaffalon, M.: The naive credal classifier. J. Stat. Plann. Infer. **105**(1), 5–21 (2002). https://doi.org/10.1016/S0378-3758(01)00201-4. Imprecise Probability Models and their Applications