Classification with decision trees from a nonparametric predictive inference perspective

Joaquín Abellán†, Rebecca M. Baker§, Frank P.A. Coolen§, Richard J. Crossman¶, Andrés. R. Masegosa†

† {jabellan,andrew}@decsai.ugr.es;
§r.m.baker@dunelm.org.uk; §frank.coolen@durham.ac.uk;
¶r.j.crossman@warwick.ac.uk
† Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
§ Department of Mathematical Sciences, Durham University, Durham, UK
¶ Warwick Medical School, University of Warwick, Coventry, UK

Summary. We present an application of Nonparametric Predictive Inference for multinomial data (NPI-M) in classification tasks.

Key words: Imprecise probabilities; Imprecise Dirichlet model; Nonparametric Predictive Inference model; uncertainty measures; supervised classification; decision trees.

1 Introduction

Many mathematical models can be used to represent the information in situations where uncertainty is present. These models are generalizations of probability theories such as belief functions, reachable probability intervals, capacities of various orders, upper and lower probabilities, and convex and closed sets of probability distributions (also called credal sets). The term *imprecise probabilities* (Klir[32], Walley [45]) subsumes these theories. Some of these generalized theories are more appropriate than others in specific situations.

With the emergence of these models, an extension of classical uncertainty-based information theory within probability theory became needed. In the 1990s, using the Shannon entropy (Shannon [43]) measure for probabilities as a starting point, a large amount of research was carried out to study measures to quantify different types of uncertainty inherent to some of these models, principally on belief functions. In recent years, this study has been extended to general credal sets. The maximum entropy measure has been used as a suitable total uncertainty measure for general credal sets[1], satisfying a number of desirable properties (Abellán et al. [6], Klir [32]).

The Imprecise Dirichlet model (IDM), presented by Walley [46], is a model for statistical inference from multinomial data which was developed to correct shortcomings of previous alternative objective models. It satisfies a set of principles which are claimed by Walley to be desirable for inference (see Walley [46]). The IDM can be seen as a model which gives imprecise probabilities that can be expressed via a set of reachable probability intervals and a belief function (Abellán [1]). The IDM has been applied to many statistical problems; a description of these applications can be seen in Bernard [10]. However, the use of the IDM has recently been questioned for some practical applications (Piatti et al. [37]). Shortcomings of the IDM were already discussed in detail by Walley [46], and by many discussants of that paper, leading Walley to strongly motivate researchers to develop alternative inference models.

Coolen and Augustin [17] presented Nonparametric Predictive Inference for multinomial data (NPI-M) as alternative, that does not suffer from some of the main drawbacks of the IDM [46]. It is different to the IDM in the sense that NPI-M learns from data in the absence of prior knowledge and with only a few modeling assumptions, most noticeably a post-data exchangeability-like assumption together with a latent variable representation of data as lines on a probability wheel. NPI-M does not satisfy some of the principles for inference suggested by Walley [46], most noticeably the Representation Invariance Principle (RPI), but Coolen and Augustin do not consider this to be a shortcoming [18]. In fact, they present arguments against general adoption of the RPI for inference and propose an alternative, weaker principle, which NPI-M satisfies.

The imprecision expressed by the NPI model is upper than the one from the IDM, when the most frequent value for the s parameter is used. It is important to remark the differences about the total uncertainty expressed by the maximum entropy function that we can find between the NPI model and the IDM. With the IDM we have more conflict but less imprecision than with the NPI model being the value of the total uncertainty of the set associated with the IDM lower than the one associated with the sets obtained from the NPI model.

One application of the study of information based uncertainty measures on imprecise probabilities is the method of Abellán and Moral [3] for building decision trees. In this method, the split criterion used has different characteristics to those of the classical split criteria. This procedure to build decision trees is able to use different models of imprecise probabilities to represent the information from a data set and also to use different uncertainty measures. The introduction of the IDM and the NPI models can produce different results for that procedure, because as we will see, the treatment of the information

[1] A measure that quantify the 2 types of uncertainty found: conflict and non-specificity

is different for each way of representation, being the imprecision greater when the NPI model is used.

We have carried out a series of experiments to check the performance of the NPI model when placed in this method and used to build decision trees. For this aim, algorithms to attain the maximum entropy probability are required; these are presented in Abellán et al. [8]. We have used 40 data sets with the common characteristic that the class variable has a known number $K \geq 3$ of cases or categories, as was considered in the model presented in Coolen and Augustin [17]. To compare results, we have used two other classical split criteria into the same scheme; a variation of the parameter used in the IDM; and others procedures of classification based on decision trees.

We will show that the model based on the NPI model has a performance with slightly improved accuracy compared to the best model based on the IDM (with a variable parameter s),[2] whilst using notably smaller trees. If the value of the parameter s in the IDM is increased, we can obtain smaller trees too but in that case, the accuracy is clearly decreased.

This paper is organized as follows: Section 2 presents a summary of the principal theories of imprecise probabilities. Section 3 describes the IDM model and Section 4 the NPI model. Section 5 is devoted to a brief overview of uncertainty measures for imprecise probabilities. In Section 6 we explain the procedure for building decision trees using imprecise probabilities and uncertainty measures, and in Section 7 we present the results of the experiments carried out. Section 8 summarizes the conclusions.

2 Theories of imprecise probabilities: A brief overview

2.1 Imprecise probabilities and credal sets

Theories of imprecise probabilities (Klir [32], Walley [45], Weichselberger [48]) share some common characteristics; for example, the evidence within each theory can be described by a lower probability function P_* on a finite finite variable X, with values in a finite set $\mathcal{X} = \{x_1, \ldots, x_K\}$, or alternatively, by an upper probability function P^* on X. These functions are always regular monotone measures (Wang and Klir [47]) and satisfy

$$\sum_{x \in X} P_*(\{x\}) \leq 1, \ \sum_{x \in X} P^*(\{x\}) \geq 1. \tag{1}$$

A general set S of imprecise probabilities on X, can be described as a set of probability distributions p on X associated with both bounds P_* and P^*, such that $p \in S \Rightarrow P_*(\{x\}) \geq p(\{x\}) \geq P^*(\{x\})$. This definition does nor force S to be closed or convex.

[2] Its normal value used is $s = 1$.

If the set of probability distributions comprises a general credal set, $\mathcal{P}$, i.e. a closed and convex set of probability distribution functions p on X (Kyburg [34]), then functions P_* and P^* associated with $\mathcal{P}$ are determined for each set $A \subseteq X$ by the expressions

$$P_*(A) = \inf_{p \in \mathcal{P}} \sum_{x \in A} p(\{x\}), \;\; P^*(A) = \sup_{p \in \mathcal{P}} \sum_{x \in A} p(\{x\}). \tag{2}$$

In this case, P_* and P^* are called dual, because for each $p \in \mathcal{P}$ and each $A \subseteq X$, the following holds:

$$P^*(A) = 1 - P_*(X - A) \tag{3}$$

where $X - A$ denotes the subset of X that is complementary to A.

2.2 Probability intervals

In the theory of probability intervals (Campos et al. [12]), the bounds $([l(x), u(x)])$ on the probability of the singleton elements $x \in X$, determine the lower and upper probabilities P_* and P^* of each event. Clearly, $l(x) = P_*(\{x\})$ and $u(x) = P^*(\{x\})$, and inequality (1) must be satisfied. Each given set of probability intervals $I = \{[l(x), u(x)] | \, x \in X\}$ is associated with a credal set, $\mathcal{P}(I)$, of probability distribution functions, p, defined as follows:

$$\mathcal{P}(I) = \{p | \, x \in X, \, p(x) \in [l(x), u(x)], \, \sum_{x \in X} p(x) = 1\}. \tag{4}$$

A given set I of probability intervals may be such that some combinations of values taken from the intervals do not correspond to any probability distribution function. This indicates that the intervals are unnecessarily wide. To avoid this deficiency, the concept of reachability was introduced by Campos et al. [12].

A given set I is called reachable if and only if for each $x \in X$ and every value $v(x) \in [l(x), u(x)]$ there exists a probability distribution function p for which $p(x) = v(x)$. The reachability of any given set I can be easily checked: the set is reachable if and only if it satisfies the following:

$$\begin{aligned} &\textstyle\sum_{x \in X} l(x) + u(y) - l(y) \leq 1, \forall y \in X, \\ &\textstyle\sum_{x \in X} u(x) + l(y) - u(y) \geq 1, \forall y \in X. \end{aligned} \tag{5}$$

The upper and lower probabilities from a reachable set of probability intervals can be obtained using the following result of Campos et al. [12]:

Proposition 1 *With the above notation, given a reachable set I of probability intervals, the lower and upper probabilities are determined for each $A \subseteq X$ by the formulae*

$$
\begin{aligned}
P_*(A) &= \max\{\textstyle\sum_{x\in A} l(x),\, 1-\sum_{x\notin A} u(x)\},\\
P^*(A) &= \min\{\textstyle\sum_{x\in A} u(x),\, 1-\sum_{x\notin A} l(x)\}.
\end{aligned}
\tag{6}
$$

3 Probability intervals from the IDM

The *imprecise Dirichlet model* (IDM) was introduced by Walley [46] for inference about the probability distribution of a categorical variable. Let us assume that X is a variable taking values on a finite set $\mathcal{X} = \{x_1, \ldots, x_K\}$ and that we have a sample of n independent and identically distributed outcomes of X. If we want to estimate the probabilities $\theta_x = p(x)$ with which X takes its values, a common Bayesian procedure consists of assuming a *prior* Dirichlet distribution for the parameter vector $(\theta_x)_{x\in\mathcal{X}}$, and then taking the *posterior* expectation of the parameters given the sample. The Dirichlet distribution depends on the parameters $s > 0$ and $\mathbf{t} = (t_x)_{x\in\mathcal{X}}$, which is a vector of positive real numbers satisfying $\sum_{x\in\mathcal{X}} t_x = 1$. The density takes the form

$$f((\theta_x)_{x\in\mathcal{X}}) = \frac{\Gamma(s)}{\prod_{x\in\mathcal{X}} \Gamma(s\cdot t_x)} \prod_{x\in\mathcal{X}} \theta_x^{s\cdot t_x - 1},$$

where Γ is the gamma function. If $n(x)$ is the number of occurrences of value x in a sample of size n, the expected *posterior* value of parameter θ_x is $\frac{n(x)+s\cdot t_x}{n+s}$, which is also the Bayesian estimate of θ_x (under quadratic loss).

The imprecise Dirichlet model [46] only depends on the parameter s, and assumes all the possible values of $\mathbf{t}$. This defines a closed and convex set of *prior* distributions. It represents a much weaker assumption than a precise *prior* model, but it is possible to make useful inferences using this model. In our particular case, where the IDM is applied to a single variable X, we obtain a credal set for this variable that can be represented by a system of probability intervals. For each parameter θ_x we obtain a probability interval given by the lower and upper *posterior* expected values of the parameter given the sample, this interval can be easily computed and is given by $[\frac{n(x)}{n+s}, \frac{n(x)+s}{n+s}]$. The associated credal set on $\mathcal{X}$ is given by all the probability distributions p' on X such that $\forall x \in \mathcal{X}$, $p'(x) \in [\frac{n(x)}{n+s}, \frac{n(x)+s}{n+s}]$. For any $A \subset \mathcal{X}$, it can be shown that

$$p'(A) \in \left[\frac{\sum_{x\in A} n(x)}{n+s}, \frac{\sum_{x\in A} n(x) + s}{n+s}\right].$$

The intervals are coherent in the sense that if they are computed by taking the infimum and supremum in the credal set, then the same set of intervals is obtained.

The parameter s determines how quickly the lower and upper probabilities converge as more data become available; larger values of s produce more cautious inferences. Walley [46] does not give a definitive recommendation, but he advocates values between $s = 1$ and $s = 2$.

Abellán [1] gives a set of properties for this type of probability interval. Every set of IDM probability intervals represents a set of reachable probability intervals. The credal set associated with a set of IDM probability intervals $\mathtt{L}$

$$\mathtt{L} = \{[l_i, u_i] |\, l_i = \frac{n(x_i)}{n+s},\, u_i = \frac{n(x_i)+s}{n+s},\, i = 1, 2, \ldots, K,\, \sum_{i=1}^{K} n(x_i) = n\},$$

can also be expressed by a belief function.

4 Probability intervals from the NPI-M

The NPI model for multinomial data (NPI-M) was developed by Coolen and Augustin [17, 18]. The model is based on a variation of Hill's assumption $A_{(n)}$ [29, 30], which relates to predictive inference involving real-valued data observations. Nonparametric predictive inference is a frequentist statistical framework with attractive properties, for which applications have been presented to many problems in statistics, reliability and operations research; for some introductions and overviews, see [9, 15, 16].

The assumption made by Coolen and Augustin whilst applying NPI for multinomial data [17, 18] is known as the circular-$A_{(n)}$ assumption, and relates to multinomial data consisting of observed values $Y_i = y_i$, $i = 1, ..., n$. These observations are related to observations of a corresponding latent variable which create n intervals on a circle; these are then represented as $I_j = (y_j, y_{j+1})$ for $j = 1, ..., n-1$, and $I_n = (y_n, y_1)$. The circular-$A_{(n)}$ assumption is that the next observation will fall into any of these intervals with equal probability $\frac{1}{n}$, so in other words $P(Y_{n+1} \in I_j) = \frac{1}{n}$ for $j = 1, ..., n$.

To present this model, we use similar notation to Coolen and Augustin [17, 18]. Suppose that there are K different categories altogether, and that the first k of these, $c_1, ..., c_k$, have already been observed. Suppose that there are n_j observations in category c_j, for $j = 1, ..., k$, and that $\sum_{j=1}^{k} n_j = n$. In this paper, we restrict attention to the case where the value of K is known.

The concept underlying NPI-M involves a latent-variable "probability wheel" representation of the data. On this probability wheel, each of our n observations is represented by a line from the center of the wheel to its boundary, such that the wheel is partitioned into n equally-sized slices. From the circular-$A_{(n)}$ assumption we conclude that the next observation has probability $\frac{1}{n}$ of being in any given slice. We must then decide which category each of these slices should represent. Coolen and Augustin [17, 18] assume that each category is only allowed to be represented by one 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. If a slice is bordered by two lines representing the same category, it must be assigned to this category. However, if a slice is bordered by two lines representing different categories, it may be assigned to any available category, that is to either of the two categories corresponding to the slice's bordering lines or to any different category not yet represented by other lines (and therefore not yet observed). It is important to emphasize that such a slice of the wheel does not have to be assigned in total to a single category, it can be divided in any way to the possible categories just mentioned.

Our general event of interest can be expressed as

$$Y_{n+1} \in \bigcup_{j \in J} c_j \tag{7}$$

where $J \subseteq \{1, ..., K\}$. We shall refer to this general event as E. Let

$$OJ = J \cap \{1, ..., k\}$$

represent the index-set for the categories in E that have already been observed, and let $r = |OJ|$. Also, let

$$UJ = J \cap \{k+1, ..., K\}$$

represent the index-set for the categories in E that have not yet been observed, and let $l = |UJ|$.

The NPI-M lower probability for the general event E (7) is found by constructing a configuration of the probability wheel which minimizes the number of slices that are assigned to E, and the NPI-M upper probability for E is found by assigning as many slices of the wheel as possible to E.

Theorem 1 *(Coolen and Augustin [18]) With the above notation, the NPI-M lower and upper probabilities for the event E based on n observations are:*

$$\underline{P}(E) = \frac{n_J - \min(K - r - l, r)}{n}$$

and

$$\overline{P}(E) = \frac{n_J + \min(r + l, k - r)}{n}$$

with $n_J = \sum_{j \in J} n_j$.

For singleton events $E = \{Y_{n+1} \in c_i\}$ we obtain the following NPI-M lower and upper probabilities:

$$\underline{P}(Y_{n+1} \in c_i) = \max\left(0, \frac{n_i - 1}{n}\right)$$

and

$$\overline{P}(Y_{n+1} \in c_i) = \min\left(\frac{n_i + 1}{n}, 1\right).$$

We consider the following set of probability intervals:

$$\mathcal{L} = \{[l_i, u_i] |\, l_i = \max\left(0, \frac{n_i - 1}{n}\right),\, u_i = \min\left(\frac{n_i + 1}{n}, 1\right),$$

$$i = 1, 2, \ldots, K,\, \sum_{i=1}^{K} n_i = n\}.$$

On this set, the following in propositions hold (Abellán et al. [8]):

Proposition 2 *$\mathcal{L}$ is a set of reachable probability intervals.*

Proposition 3 *The set of upper and lower probabilities produced by $\mathcal{L}$ is the same as the set produced by the NPI-M lower and upper probabilities of Theorem 1.*

Hence, when the NPI-M model is applied to a set of n observations, the lower and upper probabilities of an event can be obtained using only those of the singleton events. This set of lower and upper probabilities associated with the singletons expresses a reachable set of probability intervals, i.e. a credal set. An important characteristic of this credal set is that not all of its probability distributions are compatible with the theoretical NPI-M model, i.e. the set of probability distributions obtained from the NPI-M model is not a credal set. We can see this in the following example.

Example 1. (Abellán et al. [8])

Suppose we have 5 possible categories. Categories B and P are observed 4 and 5 times respectively, and the other categories R, Y and G are unobserved. The data are shown on the probability wheel below (Figure 1).

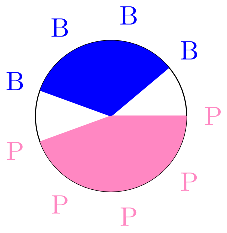

Fig. 1. Probability wheel for Example 1

Using the NPI-M lower and upper probability formulae, we find that the set of reachable probability intervals for the event $Y_{10} \in c_j$, with the outcomes ordered as $\{B, P, R, Y, G\}$, is

$$\left\{[\frac{3}{9}, \frac{5}{9}]; [\frac{4}{9}, \frac{6}{9}]; [0, \frac{1}{9}]; [0, \frac{1}{9}]; [0, \frac{1}{9}]\right\}.$$

Here, the probability distribution $D = (\frac{3}{9}, \frac{4}{9}, \frac{2}{27}, \frac{2}{27}, \frac{2}{27})$ belonging to the credal set associated with this set of reachable probability intervals. However, it is not possible to find a configuration of the probability wheel that corresponds to this probability distribution and is in line with the NPI-M framework, because of the requirement that a category may not be represented by more than one sector, as would be needed to divide the two available slices equally over the three categories R, Y and G. So we can say that the there exist probability distributions belonging to the set of all the probability distributions satisfying the NPI-M bounds which are not compatible with the NPI-M. So the set of compatible probabilities is not convex. This also implies that it is possible that we can not find a belief function associated with the set of compatible probabilities obtained from the NPI-M. This does not happen with the IDM.

Using the IDM on this example with $s = 1$, the bounds of the probability values obtained for each category, using the order $\{B, P, R, Y, G\}$, is the following:

$$\left\{ [\frac{4}{10}, \frac{5}{10}], [\frac{5}{10}, \frac{6}{10}], [0, \frac{1}{10}], [0, \frac{1}{10}], [0, \frac{1}{10}] \right\}$$

□

An approximate model can be derived from NPI-M by considering all probability distributions compatible with the set of lower and upper probabilities obtained from NPI-M [8]. This model, denoted by A-NPI-M, uses the convex hull of the set of distributions compatible with the NPI-M, and so corresponds to the structure defined by the singleton probabilities. A-NPI-M is therefore a simplification of the exact model, allowing us to avoid considering a difficult set of constraints (for more details, see Abellán et al. [8]).

5 A brief overview of uncertainty measures

It has been well established that uncertainty in classical possibility theory can be suitably quantified by the Hartley measure (Hartley [28]). For each nonempty and finite set $A \subseteq X$ of possible alternatives, the Hartley measure, $H(A)$, is defined by the formula

$$H(A) = \log_2 |A|, \tag{8}$$

where $|A|$ denotes the cardinality of A. Since $H(A) = 1$ when $|A| = 2$, H defined by equation (8) measures uncertainty in bits. The uniqueness of H was proven on axiomatic grounds by Rényi [40]. The type of uncertainty measured by H is usually called non-specificity.

In classical probability theory, a justifiable measure of uncertainty was derived by Shannon [43]. This measure, which is usually referred to as the Shannon entropy and is denoted by S, is defined for probability distribution function p on a finite set X by the formula

$$S(p) = -\sum_{x\in X} p(x)\log_2 p(x). \tag{9}$$

Since $S(p) = 1$ when $|X| = 2$ and $p(x) = 1-p(x) = 0.5$, S defined by Equation (9) measures uncertainty in bits. However, the type of uncertainty measured by the Shannon entropy is different from the uncertainty type quantified by the Hartley measure; it is well captured by the term conflict.

When the classical uncertainty theories are generalized, both types of uncertainty co-exist. This requires the Hartley measure and the Shannon entropy to be properly generalized in the various theories.

In the early 1990s, the unsuccessful attempts to find a generalized Shannon entropy in the DST were replaced with attempts to find an aggregated measure of both types of uncertainty (Harmanec and Klir [26]). An aggregate measure that satisfies all the required properties (additivity, subadditivity, monotonicity, proper range, etc.) was eventually found around the mid-1990s by several authors (see Klir [32] for more details). This aggregate uncertainty measure is a functional S^* that for each belief function Bel in the DST is defined as follows:

$$S^*(Bel) = \max_{P_{Bel}}\{-\sum_{x\in X} p(x)\log_2 p(x)\}, \tag{10}$$

where the maximum is taken over the set P_{Bel} of all probability distribution functions p that dominate the given function Bel (i.e. $Bel(A) \leq \sum_{x\in A} p(x)$ for all $A \subseteq X$). This functional can be readily generalized to any given convex set of probability distributions, as shown by Abellán and Moral (2003a).

In Abellán et al. [6] we can see that has sense to consider the following expression of the maximum entropy function:

$$S^* = (S^* - S_*) + S_*,$$

where S_* expresses the minimum entropy function. Here $S^* - S_*$ is considered as a non-specificity measure; and S_* as a conflict measure [3]

To use the maximum entropy in applications it is important to consider its calculus. Useful algorithms for computing S^* were developed for the DST by Harmanec et al. [27], for reachable interval-valued probability distributions by Abellán and Moral [2], and for the theory based on Choquet order-2 capacities (2-monotone measures) by Abellán and Moral [4].

To obtain the maximum entropy for the case of probability intervals from IDM with a value of s between 1 and 2, we can use a more efficient algorithm

[3] A total uncertainty measure, on credal sets, is composed of a part to quantify the non-specificity and a part to quantify the conflict [6]

presented in Abellán [1]. The case of probability intervals from the NPI-M and the A-NPI-M are considered using different algorithms in Abellán et al. [8], because they can express different sets of probabilities (see Example 1).

Really the value of maximum entropy function obtained for the NPI-M is very close to the one obtained for the A-NPI-M. If we consider one type of model or other, we can obtain different bounds of probabilities, as we see in the Example 1. Also, as we can see in that example, the different models used to represent the information, can express different values of uncertainty-based information. In the Example 1 the imprecision expressed by the NPI-based models is upper than the one from the IDM. It is important to remark the differences about the total uncertainty expressed by the maximum entropy function that we can find between the NPI-based models and the IDM. For example, the S^* value for the A-NPI-M is 1.3046 and the one for the IDM is 1.05319; the S_* value for the A-NPI-M is 0.6365 and the one for the IDM is 0.6730. With the IDM we have more conflict (S_*) but less non-specificity ($S^* - S_*$) than with the A-NPI-M; being the value of the total uncertainty of the set associated with the IDM around a 25% lower than the one associated with the set obtained from the A-NPI-M.

6 Procedure to build decision trees using imprecise probabilities and uncertainty measures

A decision tree, also called a classification tree, is a simple structure that can be used as a classifier. Within a decision tree, each node represents an attribute variable (or predictive attribute or feature) and each branch represents one of the states of this variable. Each tree leaf specifies an expected value of the class variable (the variable under study). The set of data used to build the decision tree is called the training set and the set used to check the model is called the test set. When we obtain a new sample or instance of the test set, we can make a decision involving a prediction about the state of the class variable by following the path through the tree from the root until a leaf is reached, by using the sample values and the tree structure.

In Figure 2 we give an example of a classification tree, involving three attribute variables A_i ($i = 1, 2, 3$), with two possible values $(0, 1)$ for each of them, and a class variable C with cases or states c_1, c_2, c_3. The root node corresponds to the empty configuration[4] (no value for any variable). Its two children are two nodes corresponding to configurations $(A_1 = 0)$ and $(A_1 = 1)$ respectively. The leaf labeled with c_3 corresponds to the configuration $\sigma = (A_1 = 1, A_3 = 0)$. In each leaf of this tree we have a single value of the class variable.

[4] The set obtained taking one state of some attribute variable is called a configuration [3]. Hence, each node (even a leaf node) in a tree determines a configuration, taking the set of states of the attribute variables from the root node to that node.

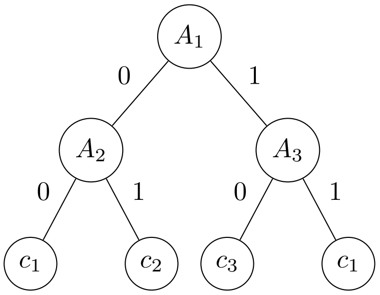

Fig. 2. Example of a classification tree.

Associated to each node is the most informative variable which has not already been selected in the path from the root to this node, as long as this variable provides more information than if it had not been included. In the latter case, a leaf node is added with the most probable class value for the partition of the data set defined with the configuration given by the path until the tree root. If two or more categories share the maximum probability for class value, we choose the most likely of these categories according to the configuration at the parent node, and so on iteratively.

One important reference for the theory of decision trees is Quinlan's ID3 algorithm [38], where precise probabilities and classical uncertainty measures on probability theory are used. Our proposal uses imprecise probabilities and uncertainty measures on more general theories than classical probability theory. We will use a similar method to the one used for the ID3 algorithm for building decision trees from a data set.

Given a data set $\mathcal{D}$, each node of a decision tree can define a set of probabilities for the class variable C in the following way: we first consider the configuration σ associated to it, and then calculate the probability intervals obtained from the IDM, NPI-M or A-NPI-M methods. We denote this set of probabilities $\mathcal{P}^{\sigma}$. For example, we have seen in Figure 2 that the node with label c_3 determines a configuration $\sigma = (A_1 = 1, A_3 = 0)$. This configuration has an associated data set, $\mathcal{D}[\sigma]$, which is the subset of the original $\mathcal{D}$ given by those cases for which $A_1 = 1$ and $A_3 = 0$. $\mathcal{P}^{\sigma}$ is the set of probabilities obtained by applying a model of imprecise probabilities.

The method starts with a tree with a single node. We shall describe it using a recursive algorithm, which starts with the empty node (the root node) with no label attached to it. Each node will have a list $\mathcal{L}^*$ of possible labels of attribute variables which can be attached to it. The procedure will start with the complete list of attribute variables.

We consider the following function :

$$\texttt{Inf}(\sigma, A_i) = \left(\sum_{a_i \in \mathcal{A}_i} r^{\sigma}_{a_i} TU(\mathcal{P}^{\sigma \cup (A_i = a_i)}) \right)$$

where $r^{\sigma}_{a_i}$ is the relative frequency with which A_i takes the value a_i in $\mathcal{D}[\sigma]$; $\sigma \cup (A_i = a_i)$ is the result of adding the value $A_i = a_i$ to configuration σ, and TU is a total uncertainty measure function, normally defined on credal sets (see Klir [32]). In the original procedure (Abellán and Moral [3]) a combined function is used, which separately measures randomness and non-specificity, or the maximum of the entropy function (Abellán and Moral [5]) which is an aggregate function of both parts of uncertainty.

If No is a node and σ the associated configuration, Inf tries to measure the weighted average total uncertainty of the sets of probabilities associated with the children of this node if variable A_i is added to it (there is a child for each one of the possible values of this node). The average is weighted by the relative frequency of each one of the children in the data set.

We now describe the method. The basic idea is very simple and is applied recursively to each of the nodes we obtain. For each of these nodes, we consider whether the total uncertainty of the credal set at this node can be decreased by adding a new node. If this is the case, then we add the node which results in the maximum decrease of uncertainty. If the uncertainty cannot be decreased, then this node is not expanded and it is transformed into a leaf of the resulting tree. We present the algorithm for this method in Figure 3.

```
Procedure BuildTree($No$,$\mathcal{L}^*$)

1. If $\mathcal{L}^* = \emptyset$, then Exit.
2. Let $\sigma$ be the configuration associated with node $No$
3. Compute the set of probabilities associated with $\sigma$
   and compute its total uncertainty $TU(\mathcal{P}^\sigma)$.
4. If $TU(\mathcal{P}^\sigma) = 0$, then Exit.
5. If $TU(\mathcal{P}^\sigma) > 0$, compute the value
         $\alpha = \min_{A_i \in \mathcal{L}^*}$ Inf$(\sigma, A_i)$
6. If $\alpha \geq TU(\mathcal{P}^\sigma)$, then Exit
7. If $\alpha < TU(\mathcal{P}^\sigma)$, then
   8. Let $A_l$ be the variable for which the minimum $\alpha$ is attained
9. Remove $A_l$ from $\mathcal{L}^*$
10. Assign $A_l$ to node $No$
11. For each possible value $a_l$ of $A_l$
   12. Add a node $No_l$
   13. Make $No_l$ a child of $No$
   14. Call BuildTree($No_l$,$\mathcal{L}^*$)
```

Fig. 3. Procedure to build decision trees using imprecise probabilities and uncertainty measures.

In this algorithm, A_l is the branching variable of node No. The intuitive idea is that when we assign this variable to No, we divide the database associated with this node among its different children. In each one of the children, we can have more precise average knowledge about C but based on a smaller sample.

The ID3 algorithm uses also uses estimates of probabilities and the uncertainty measure is again Shannon entropy. The quantity that is used to decide what variable to attach to a node is called *information gain*, which is similar to $TU(\mathcal{P}^\sigma) - \mathtt{Inf}(\sigma, A_i)$, which is what we compute to decide the branching variable. The only difference is that information gain is applied to precise probabilities. If P^σ denotes estimates of precise probabilities about C in $\mathcal{D}[\sigma]$ then the information gain is given by

$$\mathtt{InfGain}(\sigma, A_i) = S(P^\sigma) - \left(\sum_{a_i \in \mathcal{A}_i} r^\sigma_{a_i} S(P^{\sigma \cup (A_i = a_i)}) \right).$$

Note that this function can be considered as a comparison of the maximum entropy for configuration σ with the expected maximum entropy following the extension of configuration σ using the attribute variable A_i. We therefore choose to split only if the expected maximum entropy following the split is strictly less than the current maximum entropy.

The information gain has an important characteristic: it is always a non-negative number. It is important to remark that the equivalent criterion that we use in the procedure to build decision trees,

$$\mathtt{ImpInfGain} = TU(\mathcal{P}^\sigma) - \left(\sum_{a_i \in \mathcal{A}_i} r^\sigma_{a_i} TU(\mathcal{P}^{\sigma \cup (A_i = a_i)}) \right)$$

can be negative [3].

Another split criterion, used in a posterior algorithm of Quinlan [39], is the *information gain ratio*, which uses the same split criterion as the ID3 method, but splits by the entropy of the attribute variable, i.e. there exists a direct relation between this value and the number of cases of each attribute variable. This criterion penalizes variables with many states and forms the basis for the C4.5 model [39], a more complex model (defined to work with continuous variables, it works with missing data, and it has a posterior pruning process that is introduced to improve the results)

As we see in our algorithm defined above, different models based on imprecise probabilities can be applied to obtain the sets of probabilities P^σ; and different uncertainty measures TU can be applied on the set of probabilities $\mathcal{P}^\sigma$. It both these senses, then, the procedure is an open one.

7 Experiments

In this section, we will principally compare the use of the IDM, varying its parameter s, with the use of the NPI-M for building simple decision trees. In all the procedures, we will use the maximum entropy as a tool, and use splitting procedure described in the last section. The difference between the original algorithm of ID3 (Quinlan [38]), based on precise probabilities and the entropy function, and the method used here is that we use imprecise probabilities to obtain a credal set, via the IDM; or a (possible non-convex) set of probabilities, via the NPI-M, at each node; and then we use the maximum entropy measure to select the variable on which we split. For our experiments, we have used Weka software (Witten and Frank [51]) on Java 1.5, and we have added the necessary methods to build decision trees using any one of the above described split criteria.

To compare these methods, we have experimented also with the the very used and successful C4.5 algorithm of Quinlan [39]. We use the version implemented on Weka, where the C4.5 method is called J4.8 and it has a set of fixed parameters to improve the accuracy of the method. To compare results we have used this method with its parameters by defect and without a post-pruning procedure.

We have denoted the procedures using IDM with $s = 1, 1.5, 2$ as IDMs (IDM1, IDM1.5, IDM2). The rest are called as $NPI-M$ and $C4.5$ to simplify.

All five of these methods have been applied on a wide and varied collection of 40 data sets, obtained from the *UCI repository of machine learning databases*[5]. The data sets chosen are very different in terms of their sample size, number and type of attribute variables, number of states of the class variable, etc. These data have the common characteristic that their class variable has 3 or more possible values ($K \geq 3$).

A brief description of these data sets can be found in Table 1, where column "N" is the number of instances in the data sets, column "Attrib" is the number of attribute variables, column "Num" is the number of numerical variables, column "Nom" is the number of nominal (categorical) variables, column "K" is the number of cases or states of the class variable (always a nominal variable and $K \geq 3$) and column "Ka" is the number of values that the attribute variables of each data set can take.

We have applied the following preprocessing methods: missing values have been replaced with mean values (for continuous variables) and modal values (for discrete variables) using Weka's own filters. In the same way, continuous variables have been discretized using Fayyad and Irani's discretization method [21]. The preprocessing methods have been applied using the training set; the resulting values are then translated to the test set (for example, missing values in the test set are replaced by the averages or modal values obtained from the training set). For each database, we repeat 10 times a k-10 fold cross-validation

[5] They can be downloaded directly from http://archive.ics.uci.edu/ml/

procedure (100 training sets and 100 test sets). To compare the methods, we have used the following tests with a 0.05 level of significance (see Demsar [20]).

-To compare multiple classifiers on multiple datasets:

> **Friedman's test** (Friedman [22, 23]): a non-parametric test that ranks the algorithms separately for each dataset, the best performing algorithm being assigned the rank of 1, the second best, rank 2, etc. The null hypothesis is that all the algorithms are equivalent. If the null-hypothesis is rejected, we can compare all the algorithms to each other using **Holm's test** (Holm [31]).

We have considered the above set of tests following a trade-off on the recommendations by Demsar [20] and by García and Herrera [24]. As suggested in those works, the Friedman test may report a significant difference but the post-hoc test fails to detect it. This is due to the lower power of the post-hoc test conducted. Nemenyi's test [36] is recommended by Demsar, but in some situations can be a less sensitive test than others, as it is described by García and Herrera. With Nemenyi's test we can encounter situations where the differences expressed by the Friedman test were not detected. Hence, as the latter authors recommended, we considered conducting a post-hoc Holm's test.

7.1 Results

Table 2 presents the results of the accuracy of the procedures.

The Friedman test about the accuracy indicates significant differences at the 0.05 level of significance (p-value is 0.01993), and the Friedman ranks are:

IDM1:	2.625
NPI-M:	2.6125
C4.5:	3.45
IDM1.5:	2.85
IDM2:	3.4625

Here we can see that the best method (better rank) is the procedure using the NPI-M, following very close for the one with IDM1. The worse are the C4.5 and the IDM2, with very similar results.

Table 3 presents the p-values obtained from the Holm tests carried out about the accuracy. We can see that only exists significant differences between the NPI-M and the IDM2 procedures (in favor of NPI-M) at 0.05 level of significance. It must be remarked that the IDM1 is very close to obtain significant differences too with respect to the IDM2.

Table 4 presents the results of the accuracy of the procedures.

The Friedman test about the size of the trees (number of nodes) indicates significant differences at the 0.05 level of significance (p-value is 6.33E-11), and the Friedman ranks are:

Table 1. Data set description.

Data set	N	Attrib	Num	Nom	K	Ka
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	-
bridges-version1	107	11	3	8	6	2-54
bridges-version2	107	11	0	11	6	2-54
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	29	2	27	4	2-14
hypothyroid	3772	30	7	23	4	2-4
iris	150	4	4	0	3	-
letter	20000	16	16	0	26	-
lung-cancer	32	57	0	57	3	2-4
lymph	146	18	3	15	4	2-8
mfeat-factors	2000	216	216	0	10	-
mfeat-fourier	2000	76	76	0	10	-
mfeat-karhunen	2000	64	64	0	10	-
mfeat-morphological	2000	6	6	0	10	-
mfeat-pixel	2000	240	0	240	10	4-6
mfeat-zernike	2000	47	47	0	10	-
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	-
postoperative-patient-data	90	8	0	8	3	3-4
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

Table 2. Percentage of accuracy of the methods.

Dataset	IDM1	NPI-M	C4.5	IDM1.5	IDM2
anneal	99.66	99.09	99.07	99.22	99.04
arrhythmia	66.49	67.86	68.04	68.01	68.15
audiology	80.40	85.04	77.58	76.38	76.11
autos	78.30	78.11	77.34	76.95	75.10
balance-scale	69.59	69.59	69.47	69.59	69.59
bridges-version1	67.76	68.73	60.12	65.01	65.06
bridges-version2	58.99	64.15	60.95	62.75	62.20
car	91.64	90.13	93.74	87.21	86.38
cmc	48.63	48.98	47.85	48.91	49.16
dermatology	93.95	93.43	94.04	93.87	94.31
ecoli	80.27	80.19	80.24	80.18	80.21
flags	58.44	59.12	56.73	58.50	57.61
hypothyroid	99.41	99.38	99.34	99.36	99.28
iris	93.53	93.40	94.20	94.27	94.27
letter	78.15	78.77	79.42	76.15	74.62
lung-cancer	49.50	41.33	41.92	44.75	42.92
lymphography	73.12	73.68	75.67	73.14	74.65
mfeat-factors	81.47	81.71	80.56	81.11	80.57
mfeat-fourier	68.64	68.90	67.72	68.33	68.08
mfeat-karhunen	72.63	73.14	71.69	72.05	71.03
mfeat-morphological	70.76	69.78	69.91	70.18	70.05
mfeat-pixel	79.96	79.99	78.42	80.00	79.30
mfeat-zernike	63.56	64.19	62.04	63.56	62.86
nursery	96.28	95.15	98.69	94.96	93.75
optdigits	78.75	78.95	78.41	78.53	78.00
page-blocks	96.27	96.08	96.58	96.02	95.89
pendigits	89.07	89.37	89.23	88.30	87.24
postoperative-patient-data	71.00	71.11	57.56	71.11	71.11
primary-tumor	38.99	39.21	40.65	39.59	38.97
segment	94.46	94.18	94.82	94.07	93.77
soybean	91.86	93.29	92.56	91.17	89.72
spectrometer	44.77	43.34	42.94	44.75	43.82
splice	92.97	93.25	92.16	93.19	93.44
sponge	94.11	94.48	91.68	94.88	94.88
tae	46.78	46.78	46.78	46.78	46.78
vehicle	69.21	69.39	68.64	69.97	69.50
vowel	77.36	75.92	79.45	73.15	69.63
waveform	74.21	73.99	71.64	74.30	74.19
’wine	92.36	92.02	91.45	93.09	93.27
zoo	95.92	95.53	93.41	96.02	95.64
Average	76.73	76.77	75.82	76.23	75.75

Table 3. P-values Table on the accuracy for $\alpha = 0.05$. Holm's procedure rejects those hypotheses that have a p-value ≤ 0.005

i	algorithms	Holm
10	IDM2 vs. NPI-M	0.005
9	NPI-M vs. C4.5	0.005556
8	IDM1 vs. IDM2	0.00625
7	IDM1 vs. C4.5	0.007143
6	IDM1.5 vs. IDM2	0.008333
5	IDM1.5 vs. C4.5	0.01
4	IDM1.5 vs. NPI-M	0.0125
3	IDM1 vs. IDM1.5	0.016667
2	IDM1 vs. NPI-M	0.025
1	IDM2 vs. C4.5	0.05

IDM1: 4.3
NPI-M 2.7
C4.5: 3.725
IDM1.5: 2.7875
IDM2 1.4875

Here we can see that the method with a lower significant number of nodes is IDM2, the one with worse accuracy. The method with worse rank here is IDM1 follows for C4.5, IDM1.5 and NPI-M in that order.

Table 5 presents the p-values obtained from the Holm tests carried out about the number of nodes. We can see that exists significant differences between the IDM2 and all the rest ones (in favor of IDM2) at 0.05 level of significance. It must be remarked that the IDM1 is notably the worse because always exist significant differences when the test "IDM1 vs. M" is carried out, with "M" any other method in this study. When IDM1.5 and NPI-M are compared with other method "M", we always obtain a p-value lower for the test "NPI-M vs. M" than for the test "IDM1.5 vs. M".

Considering the two methods with better performance on accuracy, close to that of the NPI-M: IDM1 and IDM1.5, we have carried out a decomposition of the error in Bias and Variance (see [33]). For the sake of simplicity we do not present the tables of these results here. We have obtained the following summarized results:

Bias: NPI-M wins (low bias) in 14 and loses in 22 data sets with respect to IDM1. NPI-M wins (low bias) in 27 and loses in 12 data sets with respect to IDM1.5.

Variance: NPI-M wins in 28 and loses in 10 data sets with respect to IDM1. NPI-M wins in 23 and loses in 16 data sets with respect to IDM1.5.

The results are similar between NPI-M and IDM1: NPI-M has better variance but worse bias than IDM1; and it expresses that IDM1.5 is clearly worse than NPI-M: IDM1.5 has worse variance and worse bias than NPI-M.

Table 4. Average of the number of nodes

Dataset	IDM1	NPI-M	C4.5	IDM1.5	IDM2
anneal	71.59	65.33	62.35	67.81	65.22
arrhythmia	107.64	96.05	105.82	89.83	78.07
audiology	58.03	60.11	69.57	41.95	38.82
autos	135.75	123.91	156.88	120.65	106.70
balance-scale	27.39	27.33	15.04	27.13	27.13
bridges-version1	66.42	26.07	80.50	33.34	22.61
bridges-version2	152.82	44.97	75.43	130.74	63.95
car	161.66	125.42	186.37	113.74	76.93
cmc	314.46	195.19	351.94	289.02	223.62
dermatology	55.90	46.22	44.71	55.26	52.56
ecoli	39.75	37.05	31.69	37.25	36.23
flags	588.40	82.86	149.43	522.47	423.87
hypothyroid	60.79	56.65	50.45	53.72	47.31
iris	8.13	7.72	6.74	7.54	7.07
letter	9877.43	11346.65	13889.21	6828.23	5326.82
lung-cancer	26.15	22.20	18.15	21.76	14.76
lymphography	37.55	34.07	45.29	29.13	27.50
mfeat-factors	656.05	613.61	624.51	636.46	564.42
mfeat-fourier	718.47	638.18	827.51	622.70	500.05
mfeat-karhunen	786.20	727.34	716.69	713.82	573.60
mfeat-morphological	195.50	136.19	326.08	151.61	119.57
mfeat-pixel	978.08	898.18	937.48	972.03	903.32
mfeat-zernike	720.68	658.53	875.56	656.15	555.50
nursery	354.21	291.90	926.24	254.72	172.87
optdigits	2079.70	1906.97	1939.54	1908.70	1596.93
page-blocks	247.62	194.72	388.54	187.88	158.46
pendigits	2984.86	2666.25	3917.13	2425.42	1822.55
postoperative-patient-data	1.06	1.00	36.87	1.00	1.00
primary-tumor	127.34	89.03	129.44	102.13	86.41
segment	412.08	362.96	505.66	363.05	332.48
soybean	118.54	101.30	107.85	110.70	95.90
spectrometer	402.88	472.13	680.43	298.22	247.02
splice	419.94	360.87	503.36	335.48	229.45
sponge	6.45	5.67	16.83	5.23	4.72
tae	5.62	5.30	5.06	5.66	5.88
vehicle	205.56	162.73	312.70	172.08	145.52
vowel	409.78	399.82	530.20	317.64	271.64
waveform	739.11	446.92	2193.33	460.46	292.96
’wine	24.24	21.05	20.89	16.90	11.64
zoo	22.44	22.36	17.16	22.44	22.44
Average	610.16	589.52	796.97	480.25	383.84

Table 5. P-values Table on the number of nodes for $\alpha = 0.05$. Holm's procedure rejects those hypotheses that have a p-value ≤ 0.025.

i	algorithms	Holm
10	IDM1 vs. IDM2	0.005
9	IDM2 vs. C4.5	0.005556
8	IDM1 vs. NPI-M	0.00625
7	IDM1 vs. IDM1.5	0.007143
6	IDM1.5 vs. IDM2	0.008333
5	IDM2 vs. NPI-M	0.01
4	NPI-M vs. C4.5	0.0125
3	IDM1.5 vs. C4.5	0.016667
2	IDM1 vs. C4.5	0.025
1	IDM1.5 vs. NPI-M	0.05

8 Conclusions

Acknowledgements

This work has been supported by the Spanish “Consejería de Economía, Innovación y Ciencia de la Junta de Andalucía” under Project TIC-06016.

The work of Richard Crossman was partly funded by the UK National Institute for Health Research.

References

1. J. Abellán, Uncertainty measures on probability intervals from the imprecise Dirichlet model, International Journal of General Systems, 35(5) (2006) 509–528.
2. J. Abellán and S. Moral, Maximum entropy for credal sets, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 11(5) (2003) 587–597.
3. J. Abellán and S. Moral, Building classification trees using the total uncertainty criterion, International Journal of Intelligent Systems, 18(12) (2003) 1215–1225.
4. J. Abellán, and S. Moral, An algorithm that computes the upper entropy for order-2 capacities, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 14(2) (2005) 141–154.
5. J. Abellán, J. and S. Moral, Upper entropy of credal sets. Applications to credal classification, International Journal of Approximate Reasoning, 39(2-3) (2005) 235–255.
6. J. Abellán, G.J. Klir and S. Moral, Disaggregated total uncertainty measure for credal sets, International Journal of General Systems, 35(1) (2006) 29–44.
7. J. Abellán and A. Masegosa, An ensemble method using credal decision trees, European Journal of Operational Research, 205(1) (2010) 218–226.
8. J. Abellán, R.M. Baker and F.P.A. Coolen, Maximising entropy on the nonparametric predictive inference model for multinomial data, European Journal of Operational Research, 212(1) (2011) 112–122.

9. T. Augustin and F.P.A. Coolen, Nonparametric predictive inference and interval probability, Journal of Statistical Planning and Inference 124(2) (2004) 251–272.
10. J.M. Bernard, An introduction to the imprecise Dirichlet model for multinomial data, International Journal of Approximate Reasoning, 39 (2005) 123–150.
11. L.M. de Campos and M.J. Bolaños, Characterization of fuzzy measures through probabilities, Fuzzy Sets and Systems, 31 (1989) 23–36.
12. L.M. de Campos, J.F. Huete and S. Moral, Probability intervals: a tool for uncertainty reasoning, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2 (1994) 167–196.
13. A. Chateauneuf and J.Y. Jaffray, Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion, Mathematical Social Sciences, 17 (1989) 263–283.
14. G. Choquet, Théorie des Capacités, Annales de L'Institut Fourier, 5 (1953/54) 131–292.
15. F.P.A. Coolen, On nonparametric predictive inference and objective Bayesianism, Journal of Logic, Language and Information, 15(1-2) (2006) 21–47.
16. Coolen F.P.A., 2011. Nonparametric predictive inference. In: M. Lovric (ed.), *International Encyclopedia of Statistical Science*. Springer, Berlin, pp. 968–970.
17. F.P.A. Coolen and T. Augustin, Learning from multinomial data: a nonparametric predictive alternative to the imprecise Dirichlet model, in: F.G. Cozman, R. Nau, T. Seidenfeld (Eds.), ISIPTA'05, Proceedings of the 4th International Symposium on Imprecise Probabilities and Their Applications, Pittsburg, Pennsylvania 2005, pp. 125–134.
18. F.P.A. Coolen and T. Augustin, A nonparametric predictive alternative to the Imprecise Dirichlet Model: The case of a known number of categories, International Journal of Approximate Reasoning, 50(2) (2009) 217–230.
19. A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Annals of Mathematical Statistics, 38 (1967) 325–339.
20. J. Demsar, Statistical comparison of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.
21. U.M. Fayyad and K.B. Irani, Multi-valued interval discretization of continuous-valued attributes for classification learning, Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, (1993) 1022–1027.
22. M. Friedman, The use of rank to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (1937) 675–701.
23. M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Annals of Mathematical Statistics 11 (1940) 86–92.
24. S. García and F. Herrera, An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons, Journal of Machine Learning Research 9 (2008) 2677–2694.
25. M. Grabisch, The interaction and Möbius representations of fuzzy measures on finite spaces, k-additive measures: a survey. In: Grabisch, M. et al., Fuzzy Measures and Integrals: Theory and Applications. Springer-Verlag, New York, 2000.
26. D. Harmanec and G.J. Klir, Measuring total uncertainty in Dempster-Shafer Theory: a novel approach, International Journal of General Systems, 22 (1994) 405–419.

27. D. Harmanec, G. Resconi, G.J. Klir and Y. Pan, On the computation of uncertainty measure in Dempster-Shafer theory, International Journal of General Systems, 25(2) (1996) 153–163.
28. R.V.L. Hartley, Transmission of information, The Bell Systems Technical Journal, 7 (1928) 535–563.
29. B.M. Hill, Posterior distribution of percentiles: Bayes theorem for sampling from a population, Journal of the American Statistical Association 63 (322) (1968) 677–691.
30. B.M. Hill, de Finettis theorem, induction, and A(n) or Bayesian nonparametric predictive inference (with discussion), in: Bernardo et al. (Eds.), Bayesian Statistics 3, Oxford University Press, 1988, pp. 211–241.
31. S. Holm, A Simple Sequentially Rejective Bonferroni Test Procedure, Scandinavian Journal of Statistics 6 (1979) 65–70.
32. G.J. Klir, Uncertainty and Information: Foundations of Generalized Information Theory. John Wiley, Hoboken, New Jersey, 2006.
33. R. Kohav and D. Wolpert, Bias plus variance decomposition for zero-one loss functions. In: Proceedings of the Thirteenth International Conference of Machine Learning. (1996) 275–283.
34. H.E. Kyburg, Bayesian and non-Bayesian evidential updating, Artificial Intelligence, 31 (1987) 271–293.
35. Nadeau C. and Bengio Y. (2001) *Inference for the Generalization Error*. Machine Learning.
36. P.B. Nemenyi, Distribution-free multiple comparison. PhD thesis, Princenton University, 1963.
37. A. Piatti, M. Zaffalon and F. Trojani, Limits of learning from imperfect observations under prior ignorance: the case of the imprecise Dirichlet model, in: F.G. Cozman, R. Nau, T. Seidenfeld (Eds.), ISIPTA'05, Preceeding of the 4th International Symposium on Imprecise Probabilities and Their Applications, Pittsburg, Pennsylvania 2005, pp. 176–186.
38. J.R Quinlan, Induction of decision trees, Machine Learning, 1 (1986) 81–106.
39. J.R. Quinlan, Programs for machine learning, Morgan Kaufmann series in Machine Learning, 1993.
40. A. Rényi, Probability Theory. North-Holland, Amsterdam, 1970.
41. S.L. Salzberg, On comparison classifiers: pitfalls to avoid and a recommended approach, Data Mining and Knowledge Discovery 1 (1997) 317–328.
42. G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, 1976.
43. C.E. Shannon, A mathematical theory of communication, The Bell System Technical Journal, 27 (1948) 379–423, 623–656.
44. D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC 2000.
45. P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman & Hall, London, 1991.
46. P. Walley, Inferences from multinomial data: learning about a bag of marbles, Journal of the Royal Statistical Society B, 58 (1996) 3–57.
47. Z. Wang and G.J. Klir, Fuzzy Measure Theory, Plenum Press, New York, 1992.
48. K. Weichselberger, Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervalwahrscheinlichkeit als umfassendes Konzept, Physika, Heidelberg, 2001.

49. K. Weichselberger and S. Pöhlmann, A Methodology for Uncertainty in Knowledge-Based Systems. Springer-Verlag, New York, 1990.
50. F. Wilcoxon, Individual comparison by ranking methods, Biometrics 1 (1945) 80–83.
51. I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.