A tutorial on uncertainty modeling for machine reasoning

Branko Ristic[a,*], Christopher Gilliam[a], Marion Byrne[b], Alessio Benavoli[c]

[a] *RMIT University, Australia*
[b] *Defence Science and Technology Group, Australia*
[c] *University of Limerick, Ireland*

Abstract

Increasingly we rely on machine intelligence for reasoning and decision making under uncertainty. This tutorial reviews the prevalent methods for model-based autonomous decision making based on observations and prior knowledge, primarily in the context of classification. Both observations and the knowledge-base available for reasoning are treated as being uncertain. Accordingly, the central themes of this tutorial are quantitative modeling of uncertainty, the rules required to combine such uncertain information, and the task of decision making under uncertainty. The paper covers the main approaches to uncertain knowledge representation and reasoning, in particular, the Bayesian probabilistic, the possibilistic, the approach based on belief functions and finally the imprecise probability theory. The main feature of the tutorial is that it illustrates various approaches with several testing scenarios, and provides MATLAB solutions for them as a supplementary material for an interested reader.

Keywords: Information fusion; Uncertainty; Imprecision; Model based classification; Bayesian; Random sets; Belief function theory; Possibility functions; Imprecise probability.

[*] Corresponding author: B. Ristic, RMIT University, GPO Box 2476, Melbourne VIC 3001, Austalia; tel: +61 3 9925 3302.

Email addresses: `branko.ristic@rmit.edu.au` (Branko Ristic), `christopher.gilliam@rmit.edu.au` (Christopher Gilliam), `marion.byrne@dst.defence.gov.au` (Marion Byrne), `Alessio.Benavoli@ul.ie` (Alessio Benavoli)

Preprint submitted to Information Fusion *August 6, 2019*

1. Introduction

In most everyday situations we deal with uncertainty, from weather forecast and sporting games to medical diagnosis and investment options. The goal in all these situations is to make a decision based on uncertain domain knowledge and collected observations. In doing so, increasingly we rely on machine intelligence, which requires us to somehow quantify uncertainty in order to carry out statistical inference. In this tutorial we focus on methods of mathematical modelling of uncertain information available in the form of sensory measurements[1] and other types of information such as the prior domain knowledge, human originated statements (spoken or written), or contextual information. A decision is a choice of one of the finite number of available options, and we will primarily consider it in the context of classification of objects or phenomena. In the presence of uncertainty, the most important consideration in mathematical modeling is *integrity* [1]: a model should provide an accurate representation of available (actual) knowledge.

There is no universally accepted definition of *uncertain* information [2]. We will follow the classification due to Smets [3] and divide uncertain (or imperfect) information into *imprecise* information, information affected by variability due to *random* effects and *erroneous* information.

Information is imprecise if it denotes a set of possible values. For example, let the question be "How high is the Eiffel tower?" and the domain of legal values is the set of natural numbers (the units are metres) from 100 to 600. Specific examples of imprecise information are provided by statements such as: (i) more than 250; (ii) in the interval from 300 to 350; (iii) not less than 320; (iv) either 324 or 325. The two extremes of imprecision are: the precise (true) answer (in this case 324) and the null answer (the entire domain of legal values). Imprecision is also referred to as the *epistemic* uncertainty or incompleteness [2]. It is important to note that imprecision does not compromise the integrity of modeling, as long as the truth is in the set of possible values.

Information affected by variability due to random errors, also referred to as the *aleatory* uncertain, is the information which cannot be stated with full confidence. Going back to the Eiffel tower example, suppose we used the Thales method of similar triangles and *estimated* the tower height. The resulting estimate is uncertain due to the measurement error and this type of uncertain information is typically expressed by a probability function. A specific example could

[1] Sensory measurements are typically physical quantities, expressed by numbers.

be a probability mass function (PMF) whose support is $\{321, 322, 323, 324, 325\}$ with the confidence values $0.1, 0.2, 0.4, 0.2, 0.1$, respectively. The convention is that the confidence values sum-up to one. Note that aleatory uncertain information does not compromise the integrity of modeling, as long as the true value is assigned a non-zero confidence.

Erroneous modeling in model-based reasoning under uncertainty is often referred to as the *model-mismatch* situation. Note that this type of information compromises the integrity. Although every effort is made to avoid erroneous modeling in practice, when the modeling information is difficult to obtain[2], it may be a reality that a practitioner has to deal with.

In addition to the three main categories, other kinds of uncertainty have been defined, such as the *ambiguity* and *vagueness*. Ambiguity refers to the case when the true value is believed to be in the union of subsets of the state space. For example, an ambiguous imprecise information is that the height of the Eiffel tower is either in the interval $[200, 250]$ or $[300, 400]$. This statement implicitly assumes the confidence value 1. If we introduce confidence values less than 1 to the various intervals, for example, the confidence of 0.3 to $[200, 250]$ and 0.7 to $[300, 400]$, then the information is affected by triple uncertainty, i.e. it is ambiguous, imprecise and aleatory. Vagueness is a measure of fuzzyness when dealing with imprecision. Recall that imprecise information about the height of the Eiffel tower was expressed by intervals which represent *crisp* sets. Suppose, however, that the available information is "The height of the Eiffel tower is approximately 325 meters". This information is both vague and imprecise, and could be represented with a fuzzy interval centered at 325 [4].

The tutorial reviews the prevalent methods for quantification of uncertain (imperfect) information and subsequent model-based classification. Four such methods are discussed in detail and illustrated with examples. Section 2 presents the standard and random set-based Bayesian probabilistic method. Section 3 describes the method which uses the possibility functions to represent uncertainty. Section 4 reviews the approach based on belief functions. Finally, uncertainty reasoning using imprecise probabilities is discussed in Section 5. Throughout the paper, several testing scenarios are solved by the reviewed methods for reasoning/classification under uncertainty. MATLAB solutions are available as Supplementary Material.

[2] For example in military context, or when the repeated experiments are too expensive.

2. Bayesian approaches

Two noteworthy advocates of Bayesian model-based classification are [5, 6]. In this framework, the uncertain information is quantified using the probability functions and Bayes rule is applied to make inference. A probabilistic model is a mathematical model for explaining a phenomenon (and observations): it quantifies information and uncertainty in terms of probability distributions. It is a powerful framework but it only allows us to model the known unknowns (expressed via probability distributions). For other types of unknowns (imprecision, incorrect models), the Bayesian method in its simplest form is inappropriate and must be adapted or modified in a suitable manner [7]. One noteworthy modification is provided by Mahler [8, Ch.4-8], whose approach (see Sec. 2.2) is applicable to situations where the information (priors, measurements, likelihoods) is imprecise (possibly vague) in addition to being random.

2.1. Classical Bayesian classification

Consider a discrete random variable X, referred to as a *class*, taking values from a finite discrete space

$$\mathcal{X} = \{x_1, x_2, \cdots, x_N\} \tag{1}$$

with the cardinality (number of classes) $N > 1$. We assume, unless otherwise stated, that the set $\mathcal{X}$ is exhaustive, that is, all possible classes are included in $\mathcal{X}$. This assumption (also referred to as the *closed world* assumption), is weak in the sense that all non-accounted classes can be included in $\mathcal{X}$ under an *unknown* class, thus satisfying exhaustivity. Furthermore, we assume that the elements of $\mathcal{X}$ are mutually exclusive, meaning that an object can be classified only as a *single element* of $\mathcal{X}$.

The Bayesian method uses probability distributions to characterise uncertainty. Let the probability of an event $A \subseteq \mathcal{X}$ be denoted $P(A)$. For completeness, the axioms of probability (due to Kolmogorov) are as follows: (1) $P(A) \geq 0$ for all $A \subseteq \mathcal{X}$; (2) $P(\mathcal{X}) = 1$; (3) the probability of a union of two disjoint events[3] A_1 and A_2, is given by $P(A_1 \cup A_2) = P(A_1) + P(A_2)$. A consequence is that $P(\emptyset) = 0$. The probability mass function (PMF) $p : \mathcal{X} \to [0, 1]$, corresponding to the probability measure P, is introduced via the following relationship: $P(A) = \sum_{x \in A} p(x)$. Then clearly, $\sum_{x \in \mathcal{X}} p(x) = 1$. The PMF p assigns to each $x \in \mathcal{X}$ the probability $p(x)$ that x is the true class.

[3] This property is in general valid for any countable sequence of disjoint events.

Next we introduce the space of measured object features $\mathcal{Z}$, either as a continuous or a discrete space (possibly multi-dimensional). The inference (object classification[4]) is carried out on $\mathcal{X}$, the space which is hidden (being directly unobservable) using feature measurements. The relationship between a feature and the (hidden) object classes $x \in \mathcal{X}$ is assumed known and expressed by a probabilistic model, referred to as the *likelihood function* $g(\cdot|x)$.

Classification is carried out using Bayes formula [9, 10], which provides the probability of class $x \in \mathcal{X}$ given the feature $z \in \mathcal{Z}$:

$$p(x|z) = \frac{g(z|x)\, p(x)}{\sum\limits_{x' \in \mathcal{X}} g(z|x')\, p(x')}. \tag{2}$$

We can interpret $p(x)$ in (2) as the *prior* probability of class $x \in \mathcal{X}$ (before we process measurement z). The corresponding class probability $p(x|z)$, referred to as *posterior*, is obtained by revising the prior, using (2). Note that the key role in revising the prior plays the likelihood $g(z|x)$, as the conditional probability of z given x is true. The denominator in (2), $p(z) = \sum_{x \in \mathcal{X}} g(z|x)\, p(x)$, is just a normalisation constant which ensures that posterior satisfies $\sum_{x \in \mathcal{X}} p(x|z) = 1$. For a discrete feature space $\mathcal{Z}$, the likelihood is typically represented by a *confusion* matrix or a table of conditional PMFs $g(z|x)$, for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. If the aim is to minimize the chance of misclassification, then the classifier chooses the class with the highest posterior probability [9, Sec.1.5].

Note that the Bayesian classifier requires the uncertainty inherent in prior knowledge and measurement to be expressed precisely with the prior p and the likelihood function $g(\cdot|x)$, respectively. Later we will consider theoretical frameworks for reasoning where this requirement is relaxed.

Consider now a case where a sequence, or a collection, of feature measurements $z_{1:k} \equiv z_1, z_2, \cdots, z_k$ is available for classification, with $z_j \in \mathcal{Z}$ for $j = 1, \ldots, k$. Assuming the feature measurements are conditionally independent, Bayes formula[5] can be expressed recursively for $k = 1, 2, 3, \cdots$ as:

$$p(x|z_{1:k}) = \frac{g(z_k|x)\, p(x|z_{1:k-1})}{\sum_{x' \in \mathcal{X}} g(z_k|x')\, p(x'|z_{1:k-1})} \tag{3}$$

where at $k = 1$, by convention, $p(x|z_{1:0}) \equiv p(x)$.

[4] The term *classification* is essentially a synonym for statistical estimation on a discrete state space with no system dynamics (i.e. a class is constant in time).

[5] Bayesian classifier of this type in the machine learning literature is referred to as *naive*, because of the independence assumption. This classifier is in widespread use.

Let us demonstrate the Bayesian classifier with the following test, inspired by [5].

Testing scenario 1. Let $N = 3$ with $\mathcal{X} = \{x_1, x_2, x_3\}$, and a discrete feature space of equal cardinality N, that is $\mathcal{Z} = \{\zeta_1, \zeta_2, \zeta_3\}$. The confusion matrix is given in Table 1.(a): it is symmetric, with equal diagonal elements and its rows and columns add up to 1. This type of confusion matrix is completely specified with two parameters, N and d, where d, referred to as *diagnosticity*, represents the ratio between the diagonal and a non-diagonal element of the matrix. For the confusion matrix in Table 1.(a), $d = 5$. The prior class probabilities are given by $p(x_1) = p(x_2) = 2/5$ and $p(x_3) = 1/5$.■

Table 1: Testing scenario 1: confusion matrix and posterior probability matrix

(a) Confusion matrix

$g(\zeta_i \vert x_j)$	x_1	x_2	x_3
ζ_1	5/7	1/7	1/7
ζ_2	1/7	5/7	1/7
ζ_3	1/7	1/7	5/7

(b) Posterior probability matrix

$p(x_j \vert \zeta_i)$	ζ_1	ζ_2	ζ_3
x_1	10/13	2/13	2/9
x_2	2/13	10/13	2/9
x_3	1/13	1/13	5/9

The Bayesian classifier applies (2) to compute the *posterior* probability matrix, which for Testing scenario 1 is given by Table 1.(b). It displays probabilities $p(x_j|\zeta_i)$ for $i, j = 1, 2, 3$.

MATLAB exercise. $\gg$ `script_1(5, 3, [2/5 2/5 1/5]);`

In the problem described by the Testing scenario 1, the likelihood tables accurately represent the uncertainty due to random errors, if the feature measurement generation indeed follows the specification in Table 1.(a). In order to understand how the uncertainty due to randomness plays the role in the described classification problem, let us consider a testing scenario involving Monte Carlo simulations.

Testing scenario 2. Consider again the confusion matrix in Table 1.(a). A sequence of $K = 25$ feature measurements is generated at random following the probabilistic model specified by this confusion matrix. In doing so, the *true* class is adopted to be x_2. Repeat this procedure n times to create an ensemble of n independent Monte Carlo realisations.■

For Testing scenario 2 we apply recursively equation (3), to compute class probabilities $p(x_j|z_{1:k})$, for $j = 1, 2, 3$ and $k = 1, 2, \ldots, K$. The result, obtained by averaging over $n = 5000$ Monte Carlo realisations, is shown in Fig. 1. This figure displays the posterior probabilities $p(x_1|z_{1:k})$ (red) and $p(x_2|z_{1:k})$ (blue) with the associated error bars, indicating the plus/minus one standard deviation from the mean (truncated to interval $[0, 1]$). The third posterior probability $p(x_3|z_{1:k}) = 1 - (p(x_1|z_{1:k}) + p(x_2|z_{1:k})$ is omitted for clarity. Note that at $k = 0$, the posteriors equal the prior.

MATLAB exercise. $\gg$ `script_2(5, 3, [2/5 2/5 1/5],5000);`

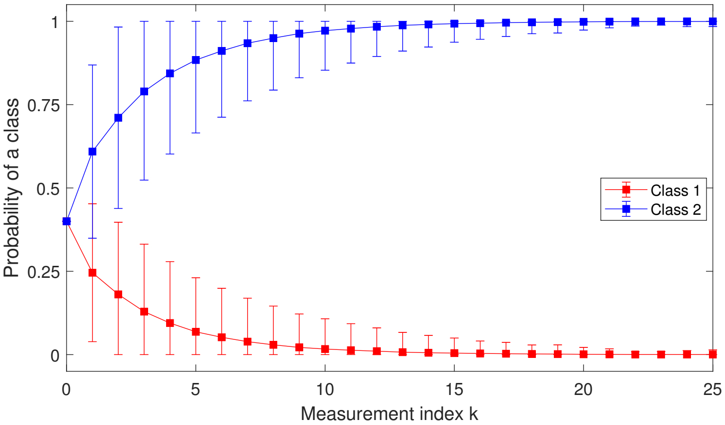

Figure 1: Bayesian classification, Monte Carlo results for Testing scenario 1: the true class is x_2. For clarity, the figure shows only the average posterior probabilities $p(x_1|z_{1:k})$ (red) and $p(x_2|z_{1:k})$ (blue), whereas $p(x_3|z_{1:k}) = 1 - p(x_1|z_{1:k}) - p(x_2|z_{1:k})$ is not shown. Error bars indicate plus/minus one standard deviation (truncated to interval $[0, 1]$).

Fig.1 indicates that, due to random errors affecting the feature measurements (characterised precisely with the confusion matrix), it takes on average $k = 9$ feature measurements for the classifier to achieve the level of confidence above 95% for the true class x_2. For a confusion matrix with a higher value of diagnosticity d, the convergence would be faster.

Multi-source combination rule. Suppose there are two classifiers. Classifier j, where $j = 1, 2$, processes a collection of measurements $z^{(j)}_{1:k_j}$ from source j. The posterior PMF resulting from classifier j is denoted by $p_j \equiv p(\cdot|z^{(j)}_{1:k_j})$. If the two sources of feature measurements are conditionally independent, the question is how to combine (fuse) the two posteriors p_1 and p_2? The fusion formula is given by [11]:

$$p(x) = \frac{p_1(x)\, p_2(x)}{\sum_{x' \in \mathcal{X}} p_1(x')\, p_2(x')}. \tag{4}$$

The key aspect of this formula is that it treats both sources as reliable and trustworthy. The formula is not applicable if the support of p_1 and the support of p_2 form two disjoint sets, because then the denominator of (4) is zero. This situation corresponds to the *total conflict* between the sources. Even if the denominator of (4) has a very small but non-zero value, the rule may be impractical, as highlighted by the so called Zadeh's example [12], where $\mathcal{X} = \{x_1, x_2, x_3\}$, with $p_1 = (0.99,\ 0.01,\ 0)$ and $p_2 = (0,\ 0.01,\ 0.99)$. Direct application of (4) to the Zadeh's example results in combined probability $p = (0,\ 1,\ 0)$, which states with 100% confidence that the truth is x_2. The result is counter intuitive, hence, as argued in [6], one has to be careful in naively applying (4). Due to a very high level of conflict between p_1 and p_2, it would make more sense to assign a certain level of confidence $0 < \alpha < 1$ to both sources, before applying (4). Thus, we should express beliefs with modified probability functions, such as $\tilde{p}_1 = (0.99\alpha,\ 0.01\alpha,\ 1-\alpha)$ and $\tilde{p}_2 = (1-\alpha,\ 0.01\alpha,\ 0.99\alpha)$. After combing $\tilde{p}_1$ with $\tilde{p}_2$ using (4), the probability mass assigned to x_1 and x_3 is given by:

$$p(x_1) = p(x_3) = \frac{0.99\alpha(1-\alpha)}{2 \cdot 0.99\alpha(1-\alpha) + 0.01^2\alpha^2}.$$

The plot of $p(x_1) = p(x_3)$ versus α is shown in Fig. 2; it reveals that even the slightest doubt in reliability of sources can create a dramatic change in the result of (4). Even for $\alpha \leq 0.998$, we can see that $p(x_1) = p(x_3) \approx 0.5$, the result which agrees with our intuition.

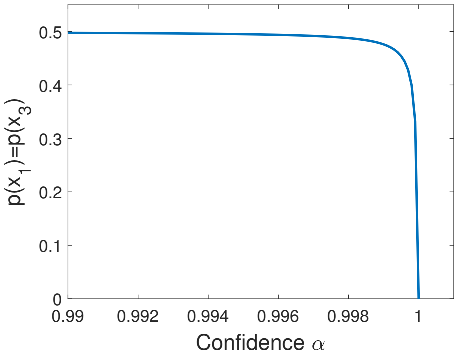

Figure 2: Zadeh example: the probability mass $p(x_1) = p(x_3)$ as a function of confidence α.

Unknown dependence. In the absence of knowledge that the two sources of measurements are conditionally independent, a variation of (4), known as the *generalised covariance intersection*, can be applied to fuse p_1 and p_2 [13, 14]:

$$p_\omega(x) = \frac{p_1^\omega(x)\, p_2^{1-\omega}(x)}{\sum_{x' \in \mathcal{X}} p_1^\omega(x')\, p_2^{1-\omega}(x')}, \tag{5}$$

where the scalar ω, which features in the exponent, satisfies $0 \leq \omega \leq 1$. The rule of combination (5) is more conservative than (4). For example, if $p_1 = p_2 = q$, the rule (5) would yield $p_\omega = q$, for any value of ω. Mahler [13] proposed to choose ω to maximise the "peakiness" of $p_\omega(x)$, i.e.

$$(\omega, x^*) = \arg\max_{\omega, x} p_\omega(x),$$

where x^* is the most probable class.

Imprecise likelihood specification. As we alluded earlier, the Bayesian approach may produce counter-intuitive results when other types of imperfect information (imprecision, incorrect models) are involved. For example, the lack of prior class probabilities is an instance of (extreme) imprecision: in this case all we know is that the entire domain of legal values (N classes) contains the truth. In the framework of Bayesian statistics, imprecision (epistemic uncertainty) is simply replaced with aleatory uncertainty via the principle of maximum entropy. This principle, when the prior is unavailable, adopts as the prior the PMF over $\mathcal{X}$ with the highest entropy, which turns out to be the uniform distribution [15]. Next we consider imprecision in specifying the relationship between the features and the classes, motivated by discussion in [16].

Testing scenario 3. Consider the space of classes $\mathcal{X} = \{x_1, x_2, x_3\}$, and assume a discrete feature space of equal cardinality, that is $\mathcal{Z} = \{\zeta_1, \zeta_2, \zeta_3\}$. The confusion matrix is not available, instead, the domain knowledge is summarised as follows[6] :
(a) class x_1 can cause observation ζ_1, only;
(b) class x_2 can cause observation ζ_1 or ζ_2;
(c) class x_3 can cause any observation from $\mathcal{Z}$.
Suppose the sequence of 17 feature measurements is available for classification. All measurements in this sequence are ζ_1, except the 7th measurement being

[6] This testing scenario could be applied to the following situation: $\mathcal{X}$ is a set of three vehicle classes: x_1 for a bus, x_2 for a passenger car and x_3 for a sports-car; $\mathcal{Z}$ is a discretised space of measured acceleration values, i.e. ζ_1 for small, ζ_2 for medium and ζ_3 for large acceleration.

$z_7 = \zeta_2$ and the 13th, being $z_{13} = \zeta_3$. The prior class probabilities are assumed equal.■

We refer to the model described above as being imprecise because there is no unique mapping from $\mathcal{X}$ to $\mathcal{Z}$ for classes x_2 and x_3. In order to use the Bayesian classifier, we are forced to adopt a confusion matrix with (precise) probability values. In the absence of further knowledge, one can invoke the principle of maximum entropy and use the uniform distribution over the subsets of $\mathcal{X}$. This results in the confusion probability matrix given by Table 2. Application of the Bayesian classifier (3) to the described sequence of 17 measurements, results in posterior class probabilities shown in Fig. 3.

Table 2: Confusion matrix for Testing scenario 3 used by the Bayesian classifier

$g(\zeta_i \mid x_j)$	x_1	x_2	x_3
ζ_1	1	1/2	1/3
ζ_2	0	1/2	1/3
ζ_3	0	0	1/3

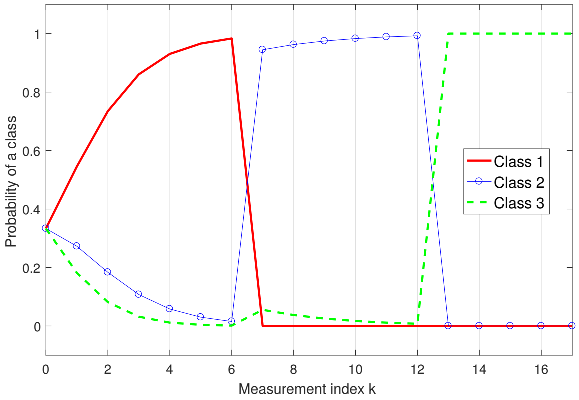

Figure 3: Testing scenario 3 - The posterior class probability versus k obtained using the Bayesian classifier. The feature sequence contains all ζ_1, except that the 7th feature is ζ_2 and the 13th feature is ζ_3.

MATLAB exercise. $\gg$ script_3;

In the Testing scenario 3, we argue that the standard Bayesian classifier gives incorrect answer until $k = 13$, when the measured feature is ζ_3. In particular, for $1 \leq k \leq 6$, when ζ_1 was repeatedly reported, the Bayesian classifier decides that the object is of class 1, although all three classes can cause this type of measurement. Similarly, for $7 \leq k \leq 12$, based on the *7th* feature measurement ζ_2, the Bayesian classifier decides that the object class is x_2, although both x_2 and x_3 can be the cause of feature ζ_2.

When we deal with imprecise likelihoods, as in the Testing scenario 3, a variant of the Bayesian classifier using random sets, performs in agreement with our intuition. This is explained next.

2.2. Mahler's approach using random sets

Mahler [8, Ch.4-8] proposed a novel approach to Bayesian estimation, fusion and classification, applicable to situations where the information (priors, measurements, likelihoods) is imprecise (possibly vague) in addition to being random. Mahler refers to this type of information as *non-traditional* information [8]. In his approach, the uncertainty inherent in non-traditional data should be represented by a special type of random variable that maps the outcome of a random experiment to a closed set, rather than to a point. Application of Mahler's approach to imprecise model-based classification was first reported in [17].

The Mahler's approach is Bayesian, and therefore uses formula (2) or (3) for inference. The difference is in the interpretation of the likelihood function $g(z|x)$. In Mahler's approach, when the measurement model is imprecise (such as in the Testing scenario 3), the likelihood function $g(z|x)$ is replaced with the *generalised* likelihood function (GLF) $\tilde{g}(z|x)$, which is defined as the probability that z belongs to the set Σ_x. The set Σ_x is the (possibly blurred) image of x on the measurement space $\mathcal{Z}$ (see Fig. 4). Thus:

$$\tilde{g}(z|x) = P(z \in \Sigma_x). \tag{6}$$

Theoretical justification for Mahler's approach can be found in [18].

Let us next illustrate Mahler's approach on Testing scenario 3. According to the statements (a) - (c) in this scenario, we can define three sets in the measurement space:

$$\begin{aligned} \Sigma_{x_1} &= \{\zeta_1\} \\ \Sigma_{x_2} &= \{\zeta_1, \zeta_2\} \\ \Sigma_{x_3} &= \{\zeta_1, \zeta_2, \zeta_3\} \end{aligned} \tag{7}$$

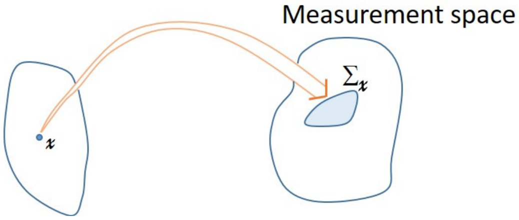

Figure 4: Imprecise mapping of $x \in \mathcal{X}$ into a set $\Sigma_x \subseteq \mathcal{Z}$.

Note from (7) that:

$$P(\zeta_1 \in \Sigma_{x_1}) = 1, \;\; P(\zeta_1 \in \Sigma_{x_2}) = 1, \;\; P(\zeta_1 \in \Sigma_{x_3}) = 1 \tag{8}$$
$$P(\zeta_2 \in \Sigma_{x_1}) = 0, \;\; P(\zeta_2 \in \Sigma_{x_2}) = 1, \;\; P(\zeta_2 \in \Sigma_{x_3}) = 1 \tag{9}$$
$$P(\zeta_3 \in \Sigma_{x_1}) = 0, \;\; P(\zeta_3 \in \Sigma_{x_2}) = 0, \;\; P(\zeta_3 \in \Sigma_{x_3}) = 1. \tag{10}$$

Then, using (6), the confusion matrix for Mahler's classifier is shown in Table 3. Next we apply Bayes formula (3), where $g(z|x)$ is replaced with $\tilde{g}(z|x)$, to the sequence of 17 measurements $z_{1:17}$ described in the Testing scenario 3. The resulting posterior class probabilities are shown in Fig. 5. We observe that Mahler's approach performs exactly in accordance with our intuition: for $1 \leq k \leq 6$ it gives equal probability to all three classes, because measurement ζ_1 is *uninformative* and hence does not change the prior. For $7 \leq k \leq 12$, it gives probability of 1/2 to classes x_2 and x_3, but zero probability to class x_1. Finally, the probability of class x_3 becomes 1 after receiving the 13th measurement.

Table 3: Confusion matrix for Testing scenario 3 used by the Mahler's classifier

$\tilde{g}(\zeta_i\|x_j)$	x_1	x_2	x_3
ζ_1	1	1	1
ζ_2	0	1	1
ζ_3	0	0	1

MATLAB exercise. $\gg$ script_4;

In summary, Mahler's approach enables Bayes formula to deal correctly with imprecise models (without violating their integrity). In the absence of imprecision in the specification of the likelihoods, such as in the Testing scenarios 1

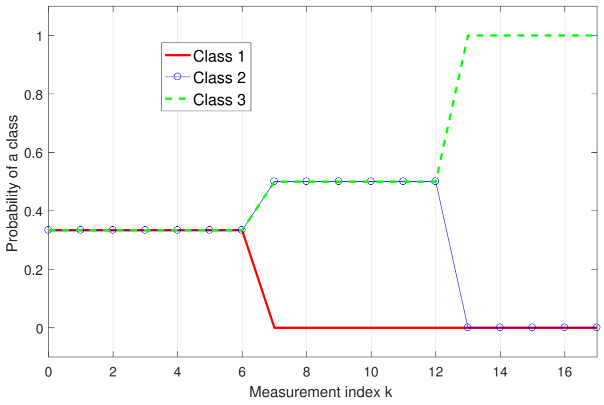

Figure 5: Testing scenario 3 - The posterior class probability versus k obtained using the Mahler approach. The feature sequence contains all ζ_1, except that the 7th feature is ζ_2 and the 13th feature is ζ_3.

and 2, Mahler's approach reduces to the classical Bayesian classifier and hence produces identical outputs.

3. Classification using possibility distributions

The theory of possibility was introduced by Zadeh [19] as an extension of his theory of fuzzy sets and fuzzy logic. Later contributions were mainly due to Dubois and Prade [20, 21, 22] and their co-workers. Possibility theory is driven by the principle of minimal specificity, meaning that any hypothesis (i.e. object class) not known to be impossible, cannot be ruled out.

For the space of object classes $\mathcal{X}$ of (1), the possibility measure of event $A \subseteq \mathcal{X}$ is a mapping $\Pi : 2^{\mathcal{X}} \rightarrow [0,1]$, where $2^{\mathcal{X}}$ is the set of all subsets of $\mathcal{X}$ (the power set). Mapping Π must satisfy three axioms: (1) $\Pi(\emptyset) = 0$; (2) $\Pi(\mathcal{X}) = 1$ and (3) the possibility of a union of disjoint events[7] A_1 and A_2 is given by $\Pi(A_1 \cup A_2) = \max[\Pi(A_1), \Pi(A_2)]$. Axiom (1) means that $\mathcal{X}$ is an exhaustive set of possible classes. The interpretation of axiom (2) is that Π is free of contradiction. Axiom (3) replaces the additivity axiom in probabilities[8].

The possibility (mass) function $\pi : \mathcal{X} \rightarrow [0,1]$ corresponding to Π is introduced via $\Pi(A) = \max_{x\in A}\pi(x)$, for every$A \subseteq \mathcal{X}$. Then clearly $\max_{x\in\mathcal{X}}\pi(x) = 1$.

[7] This property is valid for any countable sequence of disjoint events.

[8] The possibility measure belongs to a class of non-additive probabilities, see [23].

Possibility functions can represent the two extremes of imprecision as follows:

- Complete knowledge (when only one class is possible): for some x_0, $\pi(x_0) = 1$ and for all $x \in \mathcal{X}$ such that $x \neq x_0$, we have $\pi(x) = 0$.
- Complete ignorance (when all classes are possible): $\pi(x) = 1$ for all $x \in \mathcal{X}$.

We have introduced the notion of *possibility* of an event $A \subseteq \mathcal{X}$ as $\Pi(A)$. Note that a dual notion of *necessity* can be also defined as: $C(A) = \min_{x \notin A}[1 - \pi(x)]$. Duality of possibility and necessity can be expressed by $C(A) = 1 - \Pi(A^c)$, where A^c is the complement of A in $\mathcal{X}$. Axioms (1) and (2) apply to necessity, i.e. $C(\emptyset) = 0$ and $C(\mathcal{X}) = 1$, respectively. Axiom (3) takes the form: $C(A \cap B) = \min[C(A), C(B)]$. Note that possibility function π induces both Π and C (i.e. knowing π is sufficient to calculate both Π and C). One can interpret the pair necessity/possibility $[C, \Pi]$ as the *lower* and *upper* probabilities in the sense of Walley [24, 22] (to be discussed in Sec. 5).

Any probability function $p(x)$ can be turned into a possibility function $\pi(x)$, and conversely, any possibility function can be transformed into a probability function. The simplest transformations are:

$$\begin{aligned} \pi(x) &= p(x) / \max_{x' \in \mathcal{X}} p(x') && (11) \\ p(x) &= \pi(x) / \sum_{x' \in \mathcal{X}} \pi(x'). && (12) \end{aligned}$$

for all $x \in \mathcal{X}$. Other types of transformations have been also been proposed, see [22].

Classification must be carried out using a formula which relates the space of classes $\mathcal{X}$ with the observation space $\mathcal{Z}$ in a manner that is analogous to Bayes rule. Bayes-like formula to be used for this purpose is given by [25, 26]:

$$\pi(x|z) = \frac{\gamma(z|x)\,\pi(x)}{\max\limits_{x' \in \mathcal{X}} \gamma(z|x')\,\pi(x')}. \tag{13}$$

We can interpret $\pi(x)$ in (13) as the prior possibility that class $x \in \mathcal{X}$ is true. Furthermore, $\gamma(z|x)$ is the possibility of receiving measurement z given that[9] the true class is x. We refer to $\gamma(\cdot|x)$ as to the likelihood possibility function, being the analog of $g(\cdot|x)$ in (2). The denominator in (13) represents the possibility of $z \in \mathcal{Z}$, i.e. $\pi(z) = \max\limits_{x \in \mathcal{X}} \gamma(z|x)\,\pi(x)$. It ensures that the posterior possibility function satisfies $\max_{x \in \mathcal{X}} \pi(x|z) = 1$. Equation (13) is the analogue of Bayes

[9] Notion of conditioning in possibility theory is discussed in more detail in [22].

rule (2). The difference is twofold: the sum is replaced with the maximum; probability functions are replaced with possibility functions.

Making a decision, such as choosing a class based on the posterior possibility mass function $\pi(\cdot|z)$, is far from straightforward, see [22]. One simple approach, which we adopt here, is to transform the posterior possibility $\pi(\cdot|z)$ into probability via (12), followed by the selection of the class with the highest probability mass.

Next we demonstrate the performance of the possibilistic classifier with the Testing scenario 1.

Testing scenario 1. The first step is to convert the confusion matrix in Table 1.(a), which expresses the probability functions, into the corresponding confusion matrix with the possibility values. The result, obtained by applying (11) to each column of the matrix in Table 1.(a), is shown in Table 4.(a). Similarly, using (11) , the prior possibility function is $\pi(x_1) = \pi(x_2) = 1$ and $\pi(x_3) = 1/2$. Next, application of (13) results in the posterior possibility matrix shown in Table 4.(b). Finally, application of (12) to each column of the matrix in Table 4.(b) results in the corresponding posterior probability matrix. Remarkably, the result is exactly the same as shown in Table 1.(b), i.e. the possibilistic classifier and the the Bayesian classifier perform identically on the Testing scenario 1. In general, it is straightforward to prove the following property.

Property.. Formula (13) followed by conversion (12) is equivalent to Bayes rule (2) if $\gamma(z|x) = g(z|x)/\max_{z\in\mathcal{Z}} g(z|x)$ and $\pi(x) = p(x)/\max_{x\in\mathcal{X}} p(x)$.

Table 4: Testing scenario 1 - tables with possibility values

(a) Confusion possibility matrix

$\gamma(\zeta_i\|x_j)$	x_1	x_2	x_3
ζ_1	1	1/5	1/5
ζ_2	1/5	1	1/5
ζ_3	1/5	1/5	1

(b) Posterior possibility matrix

$\pi(x_j\|\zeta_i)$	ζ_1	ζ_2	ζ_3
x_1	1	1/5	2/5
x_2	1/5	1	2/5
x_3	1/10	1/10	1

MATLAB exercise. $\gg$ `script_5(5, 3, [2/5 2/5 1/5]);`

A recursive variant of (13), when a sequence or a collection of conditionally independent feature measurements $z_{1:k}$ is available for classification, is given by:

$$\pi(x|z_{1:k}) = \frac{\gamma(z_k|x)\,\pi(x|z_{1:k-1})}{\max\limits_{x'\in\mathcal{X}} \gamma(z_k|x')\,\pi(x'|z_{1:k-1})} \tag{14}$$

where by convention $\pi(x|z_{1:0}) \equiv \pi(x)$.

Next we demonstrate the performance of the possibilistic classifier with Testing scenarios 2 and 3.

Testing scenario 2. On each run a sequence of $K = 25$ feature measurements is generated following the specification in Table 1.(a), with the true class being x_2. The result obtained by averaging over 5000 runs of the Possibilistic classifier is shown in Fig. 6. This figure displays the average posterior possibility of class x_1, that is $\pi(x_1|z_{1:k})$ (red squares) and the average posterior possibility of class x_2, that is $\pi(x_2|z_{1:k})$ (blue squares). Associated error bars, indicating plus/minus one standard deviation from the mean are also shown (truncated to interval $[0, 1]$). When the possibility functions in Fig. 6 are transformed into probability functions using (12), in accordance with the previous comments, the result is exactly the same as shown in Fig. 1.

MATLAB exercise. $\gg$ script_6(5, 3, [2/5 2/5 1/5],5000);

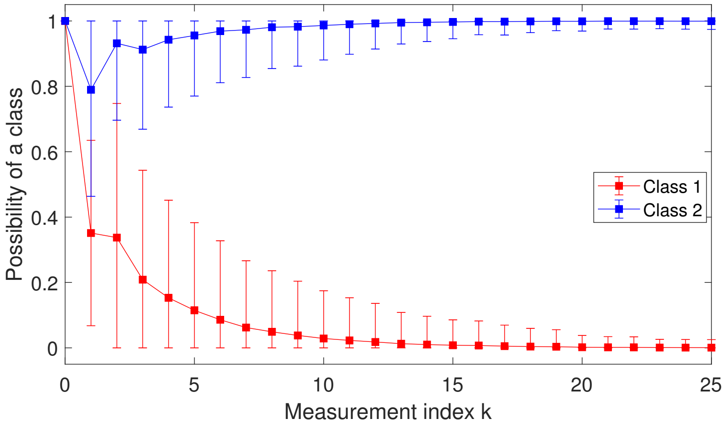

Figure 6: Possibilsitic classifier, Monte Carlo results for Testing scenario 2 (the true class is x_2): the figure shows the average posterior possibility of class x_1 and x_2 (with error bars indicating plus/minus one standard deviation, truncated to $[0, 1]$)).

Multi-source combination rule. Suppose two classifiers produce two posterior possibility functions, π_1 and π_2, respectively. Assuming the sources of feature measurements used by the two classifiers are conditionally independent, the combination rule is given by [27, Def.4]:

$$\pi(x) = \frac{\pi_1(x)\,\pi_2(x)}{\max\limits_{x' \in \mathcal{X}}[\pi_1(x')\,\pi_2(x')]}. \tag{15}$$

Eq. (15), which is the analog of (4), makes sense only if $\max_{\mathcal{X}}[\pi_1(x)\,\pi_2(x)] \neq 0$. By revisiting the Zadeh example using the possibilistic approach and the combination rule (15), one can observe that it gives the same result as the Bayesian approach. However, in dealing with imprecise likelihoods, such as in the Testing Scenario 3, the possibilistic classifier performs in accordance with our intuition.

Testing scenario 3. According to the statements (a) - (c) in this scenario, the confusion possibility matrix is given by Table 5 (which, interestingly, equates Table 3). Application of the recursive formula (14) to the sequence of 17 measurements $z_{1:17}$ described in the Testing scenario 3, results in the posterior class possibility $\pi(x|z_{1:k})$, shown in Fig. 7 versus the index k. Conversion of possibility to probability via (12), results in the posterior class probabilities identical to those in Fig. 5. Thus we observe that the possibilistic classifier performs in the same manner as the Mahler's approach, that is in agreement with our intuition.

Table 5: Confusion possibility matrix for Testing scenario 3

$\gamma(\zeta_i\|x_j)$	x_1	x_2	x_3
ζ_1	1	1	1
ζ_2	0	1	1
ζ_3	0	0	1

MATLAB exercise. ≫ script_7;

Next we introduce a testing scenario which involves a database with slightly erroneous model. In this scenario, the confusion matrix that was used in the generation of feature measurements is slightly *different* from the confusion matrix available for classification. This is a realistic scenario, because it is rarely possible to know exactly the probabilistic models in practice. Thus, the testing

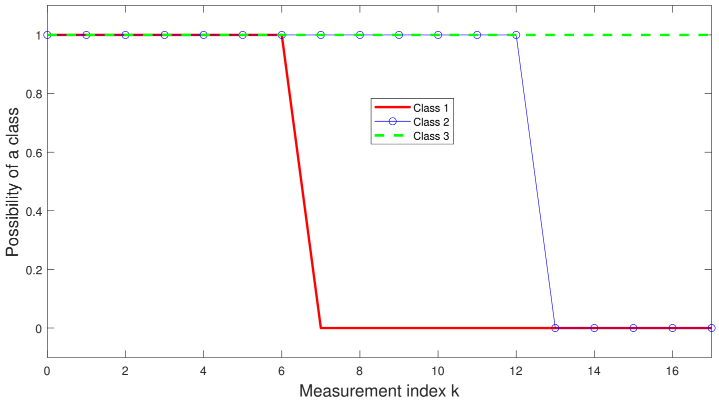

Figure 7: The posterior possibility of class 1, 2, and 3, versus k for Testing scenario 3. The input feature sequence contains all ζ_1, except that the 7th and 13th feature are ζ_2 and ζ_3, respectively.

scenario involves a compromised integrity of the database and hence evaluates the robustness of a classifier against the model-mismatch.

Testing scenario 4. Again let $N = 3$ with $\mathcal{X} = \{x_1, x_2, x_3\}$, and assume a discrete feature space of equal cardinality, that is $\mathcal{Z} = \{\zeta_1, \zeta_2, \zeta_3\}$. The confusion matrix used in the generation of feature measurements is the same as in Table 1.(a). The confusion matrix available for classification, however, is different and given by Table 6. The prior class probabilities are the same as in the Testing scenario 1, that is, $p(x_1) = p(x_2) = 2/5$ and $p(x_3) = 1/5$. The true class is x_2.■

$g(\zeta_i\|x_j)$	x_1	x_2	x_3
ζ_1	0.42	0.375	0.15
ζ_2	0.18	0.400	0.30
ζ_3	0.40	0.225	0.55

Table 6: Confusion matrix available for the classifiers in the Testing scenario 4

The results comparing the Bayesian and possibilistic classifier are shown in in Fig. 8. This figure displays the average probability of class x_2 versus the measurement index k, obtained from 5000 Monte Carlo runs. Three cases are shown: (i) the case where the correct confusion matrix is used (blue line, the same as in Fig. 1, and identical for both the Bayesian and possibilistic classifier), as an indication of the best achievable performance; (ii) the output of the

Bayesian classifier using the incorrect confusion matrix of Table 6 (green line); (iii) the output of the possibilistic classifiers when incorrect confusion matrix of Table 6 is used (red dashed line). Observe that the possibilistic classifier converges much quicker than the Bayesian classifier in this scenario. Changing the values in the incorrect confusion matrix, as expected, would affect the performance of both classifiers, however, the possibilistic classifier on average always converges faster than the Bayesian. This is a remarkable result: the possibilistic classifier is more robust to the model mismatch than the Bayesian classifier. An explanation for this phenomenon is that possibility functions represent only the upper bounds of the corresponding probability functions and thus are more resilient to the small errors in probabilistic models. Similar results have been observed in the context of estimation theory [28].

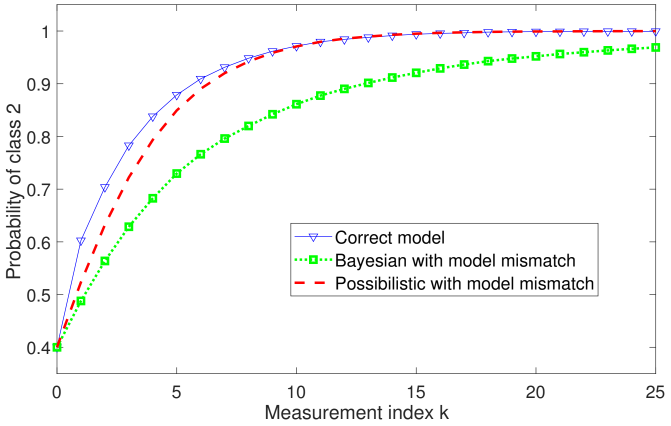

Figure 8: Results for the Testing scenario 4 (model mismatch); the true class is x_2. The figure shows the average posterior probability of class 2 vs the measurement index k, computed by the Bayesian and the possibilistic classifier.

MATLAB exercise. $\gg$ script_8(5, 3, [2/5 2/5 1/5],5000);

4. Classification using belief functions

The specification of probability distributions typically demands more information than is really available. Belief functions [29] were introduced for modeling uncertainty (about knowledge, opinions, judgments and evidence) in order to allow a more realistic and flexible treatment. In this tutorial we will mainly focus on an interpretation of the belief function theory (also known as the Dempster-Shafer (DS) theory) [29, 30] referred to the *transferrable belief model* (TBM), developed by Smets and his coworkers [31, 32, 33, 16, 34]. An early (pre-TBM) attempt to apply the DS theory to model-based classification [35] was vigorously attacked in [5] and hence did not make a lasting impact. The TBM framework, on the other hand, has become one of the most favourable approaches to object classification, primarily because: (a) it is general enough to quantify any type of uncertainty and (b) it provides a plethora of tools for object classification (e.g. uncertain implication rule, methods for dealing different granularities of the space of classes), which are not readily available in other approaches.

4.1. TBM approach

The central element of the TBM is the *basic belief assignment* (bba) $m : 2^{\mathcal{X}} \to [0,1]$, which expresses the belief over the power-set $2^{\mathcal{X}} = \{A : A \subseteq \mathcal{X}\}$, which consists of all the subsets of the space of object classes $\mathcal{X}$. A bba must satisfy:

$$\sum_{A \subseteq \mathcal{X}} m(A) = 1. \tag{16}$$

The value of $m(A)$ represents the amount of belief that is exactly committed to $A \subseteq \mathcal{X}$, and due to lack of further information, cannot be transferred to any more specific subset or element of A. Note that this is in contrast to the probability functions $p : \mathcal{X} \to [0,1]$ and possibility functions $\pi : \mathcal{X} \to [0,1]$, which both express the basic beliefs over the singletons of $\mathcal{X}$. Note also that the cardinality of the power-set grows with N as 2^N. The bba assigned to the empty set, $m(\emptyset)$, can be interpreted as the amount of conflict or the possibility that $\mathcal{X}$ is not an exhaustive set. The subsets $A \subseteq \mathcal{X}$ with the property $m(A) > 0$, are referred to as the focal sets of bba m. The state of complete ignorance is represented by a *vacuous* bba, specified as: $m(A) = 1$ if $A = \mathcal{X}$, and zero otherwise.

There are several alternative and convenient quantifications of belief, all in one-to-one correspondence with the bba m. In this overview paper we only introduce the belief function and the plausibility function.

The belief function. $bel : 2^{\mathcal{X}} \to [0,1]$ is defined for all $A \subseteq \mathcal{X}$ as

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \tag{17}$$

and represents the total belief that is committed to A, without also being committed to its complement A^c. The belief function satisfies the following conditions [29]: (i) $bel(\emptyset) = 0$; (ii) $bel(\mathcal{X}) = 1$; (iii) for any positive integer n and every collection of subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots \mathcal{X}_n \subseteq \mathcal{X}$:

$$bel\left(\bigcup_i \mathcal{X}_i\right) \geq \sum_{I \subseteq \{1,\ldots,n\}, I \neq \emptyset} (-1)^{|I|+1}\, bel\left(\bigcap_{i \in I} \mathcal{X}_i\right) \tag{18}$$

where $|I|$ denotes the cardinality of set I. Condition (iii) is referred to as ∞-monotonicity.

The plausibility function. $pl : 2^{\mathcal{X}} \to [0,1]$ is defined for all $A \subseteq \mathcal{X}$ as

$$pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{19}$$

and represents the total belief which does not contradict A.

The use of the TBM for object classification has been promoted in [33]. This reference refutes all arguments, presented in [5] about the superiority of Bayesian classification, with numerous examples. As we mentioned earlier, the TBM framework provides a very rich collection of tools for manipulation of data in order to carry out object classification, and next we briefly review some of them.

Conjunctive combination. Consider the case where two experts (or classifiers) provide their belief in the form of two bbas, defined on the same space $\mathcal{X}$. Let the two bbas be m_1 and m_2, and suppose the they originate from two distinct[10] pieces of evidence. Then the joint impact of the two pieces of evidence can be expressed by the bba $m_{12} = m_1 \bigcirc\!\!\!\!\!\cap\, m_2$ defined as [33]:

$$m_{12}(A) = \sum_{\substack{B,C \subseteq \mathcal{X} \\ B \cap C = A}} m_1(B) \cdot m_2(C), \quad \forall A \subseteq \mathcal{X} \tag{20}$$

The conjunctive rule of combination (20) is both commutative and associative. Note that combination (20) may result in a *conflict*, which is manifested by $m_{12}(\emptyset) > 0$. Providing that $m_{12}(\emptyset) < 1$, the normalised bba $\tilde{m}_{12}$ (which is free of conflict), corresponding to m_{12}, can be computed as

$$\tilde{m}_{12}(A) = \begin{cases} m_{12}(A)/(1 - m_{12}(\emptyset)), & \text{for all } A \neq \emptyset \\ 0, & \text{if } A = \emptyset. \end{cases} \tag{21}$$

[10] The notion of distinctness, discussed in [32], is similar to independence.

The belief mass given to empty set, i.e. $m_{12}(\emptyset)$, is often used in the TBM framework as a distance measure between bba's m_1 and m_2 [36, 37]. If this distance is too high (close to 1), this is an indication that the two bba's may not be compatible (and perhaps should not be fused). This could happen, for example, if m_1 and m_2 characterise two different entities (e.g. due to a data association error).

It is straightforward to show that if m_1 and m_2 are two probability functions, then the normalised conjunctive combination (21) reduces to the combination rule (4).

The generalised Bayes theorem (GBT). Recall that the Bayes formula (2) combines the beliefs expressed over *two* state spaces, the space of classes $\mathcal{X}$ and the measurement space $\mathcal{Z}$. This concept has been extended in the TBM framework as follows [33, 34]. A sensor measurement is now an element of the power set $2^{\mathcal{Z}}$. The likelihood (confusion) table is replaced with the bba table over $\mathcal{Z}$, conditioned on $x_i \in \mathcal{X}$ being the true class ($i = 1, \dots, N$). Let us denote this bba table by $m^{\mathcal{Z}}(\cdot|x_i)$, where the superscript $\mathcal{Z}$ is introduced to emphasise that this bba is defined over the space $2^{\mathcal{Z}}$. Given a measurement $B \in 2^{\mathcal{Z}}$, the GBT provides a formula for computing the conditional bba[11] $m^{\mathcal{X}}(\cdot|B)$ expressed over the space of classes $\mathcal{X}$, such that for every $A \subseteq \mathcal{X}$:

$$m^{\mathcal{X}}(A|B) = \prod_{x_i \in A} pl^{\mathcal{Z}}(B|x_i) \prod_{x_i \in A^c} [1 - pl^{\mathcal{Z}}(B|x_i)], \quad \forall A \subseteq \mathcal{X}, B \subseteq \mathcal{Z} \tag{22}$$

where $pl^{\mathcal{Z}}(\cdot|x_i)$ is the conditional plausibility function computed from $m^{\mathcal{Z}}(\cdot|x_i)$ via (19). The GBT does not incorporate the prior belief over $\mathcal{X}$, that is, the prior is by default considered to be a vacuous bba. If, however, some prior belief in the form of bba $m^{\mathcal{X}}$ is available, then it can be combined with $m^{\mathcal{X}}(\cdot|B)$, resulting from (22), using the conjunctive rule of combination (20).

Suppose a sequence of k independent feature measurements

$$B_{1:k} \equiv B_1, B_2, \cdots, B_k$$

is available for classification, where $B_j \in 2^{\mathcal{Z}}$, for $j = 1, 2, \dots, k$. The GBT can be expressed recursively for $k = 2, 3, \cdots$ as:

$$m^{\mathcal{X}}(A|B_{1:k}) = m^{\mathcal{X}}(A|B_k) \textcircled{\cap} m^{\mathcal{X}}(A|B_{1:k-1}) \tag{23}$$

[11] For more details on conditioning in TBM, see for example [33].

for all $A \subseteq \mathcal{X}$, where $m^{\mathcal{X}}(A|B_k)$ is computed using (22). The recursion starts at $k = 1$ with

$$m^{\mathcal{X}}(A|B_{1:1}) = m^{\mathcal{X}}(A|B_1) \textcircled{\cap} m^{\mathcal{X}}(A)$$

where bba $m^{\mathcal{X}}$ represents the prior belief over the space of classes $\mathcal{X}$. We will refer to $m^{\mathcal{X}}(\cdot|B_{1:k})$ as to the *posterior bba* at the measurement index k.

Decision making and the pignistic probability. Uncertainty represented by a bba $m^{\mathcal{X}}$ is not practical for making a decision on the object class. A comprehensive review on decision making using belief functions is presented in [38]. Because decision making is straightforward with probabilities, a common approach is to first map a bba to a probability function, and then, in the context of classification, to choose the class with the highest probability. Note that there are infinitely many such mappings. Smets [39] introduced the pignistic transform for this purpose, mapping a bba to a probability function referred to as the *pignistic probability* (Later we introduce another mapping, due to Voorbraak [40]). For singletons $x_i \in \mathcal{X}$, the pignistic probability is defined as

$$P_m^{\text{pign}}(x_i) = \sum_{A \subseteq \mathcal{X} \text{ } s.t. \text{ } A \ni x_i} \frac{1}{|A|} \frac{m^{\mathcal{X}}(A)}{[1 - m^{\mathcal{X}}(\emptyset)]} \tag{24}$$

The belief function and the plausibility function act as the lower and upper limit, respectively, of the pignistic probability, that is: $bel(x_i) \leq P_m^{\text{pign}}(x_i) \leq pl(x_i)$.

Now we have all the necessary tools to apply the TBM classifier to the Testing scenarios 1 and 2.

Testing scenarios 1 and 2. Assuming that the confusion matrix in Table 1.(a) is trustworthy, we can simply represent it as a bba matrix shown in Table 7 (note that a measurement in this table is an element of $2^{\mathcal{Z}}$). Furthermore, assuming that the prior class probabilities are correct, we can represent them as a bba with three focal sets: $m^{\mathcal{X}}(\{x_1\}) = m^{\mathcal{X}}(\{x_2\}) = 2/5$, $m^{\mathcal{X}}(\{x_3\}) = 1/5$. Then, application of the GBT (22) followed by the pignistic transform (24) results in the posterior probability matrix identical to the one shown in Table 1.(b).

MATLAB exercise. ≫ `script_9(5, 3, [2/5 2/5 1/5]);`

Continuing with Monte Carlo simulations (Testing scenario 2) involving a sequence of $K = 25$ feature measurements and using the recursive form of the GBT (23), we expect to obtain the classification results identical to those shown in Fig. 1. This, however, is true only if we occasionally normalise the bba $m^{\mathcal{X}}(\cdot|B_{1:k})$, using (21). For example, we can apply normalisation, after application of the GBT (23), if the belief mass given to the empty-set is higher than 0.9.

$m^{\mathcal{Z}}(\cdot\vert x_j)$	x_1	x_2	x_3
$\emptyset$	0	0	0
$\{\zeta_1\}$	5/7	1/7	1/7
$\{\zeta_2\}$	1/7	5/7	1/7
$\{\zeta_1, \zeta_2\}$	0	0	0
$\{\zeta_3\}$	1/7	1/7	5/7
$\{\zeta_1, \zeta_3\}$	0	0	0
$\{\zeta_2, \zeta_3\}$	0	0	0
$\{\zeta_1, \zeta_2, \zeta_3\}$	0	0	0

Table 7: Confusion matrix in Table 1.(a), for Testing scenarios 1 and 2, expressed with bbas

MATLAB exercise. $\gg$ script_10(5, 3, [2/5 2/5 1/5],1000);

In summary, because the probability functions are a special case of the basic belief assignments, if they are trustworthy, we can use them in the TBM framework unchanged. The GBT, followed by bba normalisation, in this case is identical to the Bayes rule. Consequently, the classification results by TBM equal those obtained using the Bayesian classifier.

Testing scenario 3. The TBM framework deals elegantly with the scenario which involves a database with imprecise feature-to-class specification. The bba specification of the confusion matrix for this scenario is given in Table 8. Application of the TBM classifier (i.e. the recursive GBT (23) followed by the pignistic transform (24)) to the sequence of 17 measurements $z_{1:17}$ using the confusion matrix in Table 8, results in the plot identical to the one shown in Fig. 5. We have already argued that this is the correct output.

MATLAB exercise. $\gg$ script_11;

The least committed bba. Before we apply the TBM framework to the Testing scenario 4, we need to introduce the concept of the *least committed* (LC) bba, corresponding to the probability function p. If there is no reason to believe that p is a trustworthy (objective) representation of uncertainty (but rather only a subjective model), then we may prefer to replace p by the most cautious bba corresponding to p [32]. This bba is denoted m^p_{LC} and is referred to as the LC bba. The pignistic transform of m^p_{LC} equals p, that is $P^{\text{pign}}_{m^p_{\text{LC}}} = p$. Note that m^p_{LC} is only one among an infinite number of iso-pignistic bbas and is

$m^{\mathcal{Z}}(\cdot\|x_j)$	x_1	x_2	x_3
$\emptyset$	0	0	0
$\{\zeta_1\}$	1	0	0
$\{\zeta_2\}$	0	0	0
$\{\zeta_1, \zeta_2\}$	0	1	0
$\{\zeta_3\}$	0	0	0
$\{\zeta_1, \zeta_3\}$	0	0	0
$\{\zeta_2, \zeta_3\}$	0	0	0
$\{\zeta_1, \zeta_2, \zeta_3\}$	0	0	1

Table 8: The confusion matrix for Testing scenario 3, expressed with bbas for TBM classification

built from p as follows [32, 16]. First, let us re-label the elements of $\mathcal{X}$ so that $p(x^{(1)}) \geq p(x^{(2)}) \geq \cdots \geq p(x^{(N)})$. The LC bba has up to N focal sets defined as $A_i = \{x^{(1)}, x^{(2)}, \ldots, x^{(i)}\}$, where $i = 1, \ldots, N$. Note that the focal sets are nested, that is $A_1 \subset A_2 \subset \cdots \subset A_N$. The belief mass allocated to each focal set A_i is given by [32]:

$$m^p_{\text{LC}}(A_i) = |A_i| \left(p(x^{(i)}) - p(x^{(i+1)})\right) \tag{25}$$

for $i = 1, \ldots, N$ and with the convention $p(x^{(N+1)}) = 0$.

Testing scenario 4. Each column in the confusion matrix in Table 6 is treated as a subjective probability function. The corresponding LC bbas then represent the columns of the confusion matrix to be used by the TBM classifier, shown in Table 9. The average probability of class x_2 versus the measurement index k, obtained from 5000 Monte Carlo runs, is shown in Fig. 9. Observe that the least-committed TBM classifier converges quicker than the Bayesian classifier, but slower than the possibilistic classifier.

MATLAB exercise. $\gg$ script_12(5, 3, [2/5 2/5 1/5],5000);

4.2. Dealing with different subspaces

The most convincing arguments, put forward by Smets and his co-workers, in favour of the TBM versus the Bayesian probabilistic approach, involve the combination of beliefs expressed at different subspaces. For example [33], suppose there are $N = 3$ classes, i.e. $\mathcal{X} = \{x_1, x_2, x_3\}$. Type 1 feature can distinguish

$m^{\mathcal{Z}}(\cdot\vert x_j)$	x_1	x_2	x_3
$\emptyset$	0	0	0
$\{\zeta_1\}$	0.02	0	0
$\{\zeta_2\}$	0	0.025	0
$\{\zeta_1, \zeta_2\}$	0	0.3	0
$\{\zeta_3\}$	0	0	0.25
$\{\zeta_1, \zeta_3\}$	0.44	0	0
$\{\zeta_2, \zeta_3\}$	0	0	0.3
$\{\zeta_1, \zeta_2, \zeta_3\}$	0.54	0.675	0.45

Table 9: Confusion matrix, built from Table 6, for the Testing scenarios 4 and TBM classification. The columns are computed as the least committed bbas corresponding to those in Table 6.

(in a probabilistic sense) between the objects of class x_1 and class x_2, while its relationship to x_3 is unknown. Similarly, type 2 feature can distinguish between x_2 and x_3, but its relationship to x_1 is unknown. Let us denote the beliefs expressed upon receiving features of type 1 and 2 with bbas $m_1^{\{x_1,x_2\}}$ and $m_2^{\{x_2,x_3\}}$, respectively. The problem is how to combine $m_1^{\{x_1,x_2\}}$ and $m_2^{\{x_2,x_3\}}$, being the two bbas on different spaces?

The TBM introduces the concept of *ballooning extension* in order to solve this problem, without making additional assumptions (i.e. without compromising the integrity of the database). In the example above, the TBM would apply the ballooning extension to represent both bbas, $m_1^{\{x_1,x_2\}}$ and $m_2^{\{x_2,x_3\}}$, on the common space $\{x_1, x_2, x_3\}$, followed by the conjunctive rule of combination (20). The ballooning extension plays the key role in the derivation of the GBT (22) and in object classification using uncertain implication rules [16],[41].

Ballooning extension. Let $\mathcal{X}$ be a space of classes and let $\mathcal{X}'$ be a subset of $\mathcal{X}$. Given a bba $m^{\mathcal{X}'}$ defined on $\mathcal{X}'$, its corresponding bba on $\mathcal{X}$, which does not compromise integrity[12], is denoted as $m^{\mathcal{X}'\Uparrow\mathcal{X}}$ and expressed as [33]:

$$m^{\mathcal{X}'\Uparrow\mathcal{X}}(A) = \begin{cases} m^{\mathcal{X}'}(A'), & \text{if } A' \subseteq \mathcal{X}', A = A' \cup (\mathcal{X}')^c \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Testing scenario 5. Let the space of object classes be $\mathcal{X} = \{x_1, x_2, x_3\}$. Two bbas with partially overlapping spaces, $m_1^{\{x_1,x_2\}}$ and $m_2^{\{x_2,x_3\}}$, are specified in

[12] The least committed bba.

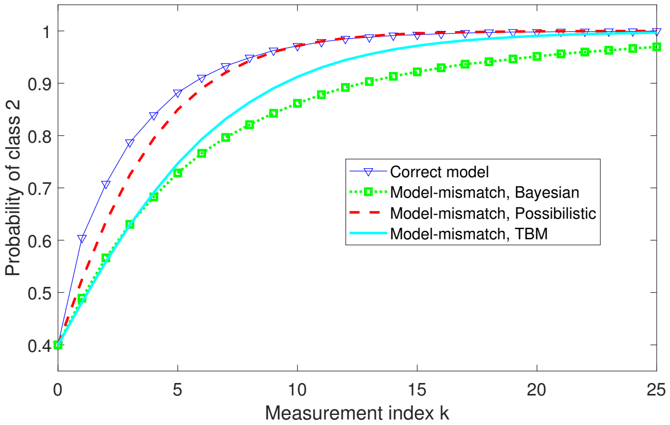

Figure 9: Testing scenario 4 (model mismatch) results for the TBM classifier: the average posterior probability of class 2 vs measurement index k (cyan solid line). The true class is x_2. The remaining curves are the same as in Fig.8.

Table 10. Note that both bbas are probability functions. How to combine these two pieces of information and determine the most probable class?■

Table 10: Testing scenario 5 involving two bbas on partially overlapping spaces

	$m_1^{\{x_1,x_2\}}$	$m_2^{\{x_2,x_3\}}$
$\{x_1\}$	0.6	
$\{x_2\}$	0.4	0.7
$\{x_1,x_2\}$		
$\{x_3\}$		0.3
$\{x_1,x_3\}$		
$\{x_2,x_3\}$		
$\{x_1,x_2,x_3\}$		

The TBM solution is given in Table 11. Columns $m_1^{\mathcal{X}}$ and $m_2^{\mathcal{X}}$ in Table 11 were obtained by application of the ballooning extension to $m_1^{\{x_1,x_2\}}$ and $m_2^{\{x_2,x_3\}}$, respectively. The column $m_1^{\mathcal{X}} \textcircled{\cap} m_2^{\mathcal{X}}$ is a result of the conjunctive rule of combination (20). Finally the two last columns in Table 11 display the pignistic probability (see (24)) and the Voorbraak probability, explained next.

Table 11: TBM solution for the Testing scenario 5

	$m_1^{\mathcal{X}}$	$m_2^{\mathcal{X}}$	$m_1^{\mathcal{X}} \textcircled{\cap} m_2^{\mathcal{X}}$	$P^{\text{pign}}_{m_1^{\mathcal{X}} \textcircled{\cap} m_2}$	$P^{\text{Voor}}_{m_1^{\mathcal{X}} \textcircled{\cap} m_2}$
$\{x_1\}$			0.42	0.51	0.51
$\{x_2\}$			0.28	0.28	0.24
$\{x_1, x_2\}$		0.7	0		
$\{x_3\}$			0.12	0.21	0.25
$\{x_1, x_3\}$	0.6	0.3	0.18		
$\{x_2, x_3\}$	0.4		0		
$\{x_1, x_2, x_3\}$			0		

Voorbraak probability. Voorbraak transform [40] is a mapping from the belief space to the probability space, defined for every $x_i \in \mathcal{X}$ as:

$$P_m^{\text{Voor}}(x_i) = \frac{\sum\limits_{A \subseteq \mathcal{X} \ s.t. \ A \ni x_i} m(A)}{\sum\limits_{B \subseteq \mathcal{X}} m(B) \cdot |B|} \tag{27}$$

The Voorbraak probability has the following remarkable property. Suppose two bba $m_1^{\mathcal{X}}$ and $m_2^{\mathcal{X}}$ are combined using the conjunctive rule (20). The Voorbraak probability of the combined bba equals the probability function obtained by application of the combination rule (4) to $P_{m_1}^{\text{Voor}}$ and $P_{m_2}^{\text{Voor}}$.

A solution to the Testing scenario 5 using the standard probability theory and the possibilistic approach can be formulated by introducing a free parameter $\beta > 0$. Probability functions $p_1^{\mathcal{X}}$ and $p_2^{\mathcal{X}}$, corresponding to Table 10, are expressed in Table 12.(a). For example, $p_1^{\mathcal{X}}$ is obtained as follows: $p_1^{\mathcal{X}}(x_1) \propto 0.6$, $p_1^{\mathcal{X}}(x_2) \propto 0.4$ and $p_1^{\mathcal{X}}(x_3) \propto \beta$. Then the normalisation constant is $0.4 + 0.6 + \beta = 1 + \beta$.

Possibility functions $\pi_1^{\mathcal{X}}$ and $\pi_2^{\mathcal{X}}$, corresponding to Table 10, are expressed in Table 12.(b). The normalisation constant for $\pi_1^{\mathcal{X}}$ is $\max[0, 4, 0.6, \beta] = \max[0.6, \beta]$. We can combine $p_1^{\mathcal{X}}$ and $p_2^{\mathcal{X}}$ using (4). Accordingly, $\pi_1^{\mathcal{X}}$ can be combined with $\pi_2^{\mathcal{X}}$ using (15). If we choose $\beta = 1$, both approaches, remarkably, yield the same result as the TBM, followed by the Voorbraak transform (the last column of Table 11).

MATLAB exercise. $\gg$ script_13;

Table 12: Testing scenario 5: probabilistic and possibilistic approach

(a) Probability functions

	$p_1^{\mathcal{X}}$	$p_2^{\mathcal{X}}$
x_1	$\frac{0.6}{\beta+1}$	$\frac{\beta}{\beta+1}$
x_2	$\frac{0.4}{\beta+1}$	$\frac{0.7}{\beta+1}$
x_3	$\frac{\beta}{\beta+1}$	$\frac{0.3}{\beta+1}$

(b) Possibility functions

	$\pi_1^{\mathcal{X}}$	$\pi_2^{\mathcal{X}}$
x_1	$\frac{0.6}{\max[\beta,0.6]}$	$\frac{\beta}{\max[\beta,0.7]}$
x_2	$\frac{0.4}{\max[\beta,0.6]}$	$\frac{0.7}{\max[\beta,0.7]}$
x_3	$\frac{\beta}{\max[\beta,0.6]}$	$\frac{0.3}{\max[\beta,0.7]}$

Uncertain implication rules. Sometimes it can be useful to express uncertain prior knowledge that relates two different spaces, e.g. $\mathcal{X}$ and $\mathcal{Y}$, in the form of uncertain implication rules. An implication rule R is formally an expression:

$$A \subseteq \mathcal{X} \Rightarrow B \subseteq \mathcal{Y}. \tag{28}$$

A shortened notation for (28) is $A \Rightarrow B$. The rule (28) can be assigned a confidence level $\alpha \in [0, 1]$. The expression for a bba which represents the uncertain implication rule (28) on the joint space $\mathcal{X} \times \mathcal{Y}$ was derived using the ballooning extension [16] by exploiting the logical equivalence between $A \Rightarrow B$ and "not A or B". This bba has two focal sets:

$$m_R^{\mathcal{X}\times\mathcal{Y}}(C) = \begin{cases} \alpha, & \text{if } C \in (A \times B) \cup (A^c \times \mathcal{Y}) \\ 1-\alpha, & \text{if } C \in \mathcal{X} \times \mathcal{Y}. \end{cases} \tag{29}$$

Note that the bba $m_R^{\mathcal{X}\times\mathcal{Y}}$ assigns the belief mass $1-\alpha$ to the joint space $\mathcal{X} \times \mathcal{Y}$, which represents the complete ignorance.

Marginalisation. Consider a bba defined on a joint space $\mathcal{X} \times \mathcal{Y} = \{(x_i, y_j); i = 1, \dots, N; j = 1, \dots, M\}$. Marginalisation is a projection of bba $m^{\mathcal{X}\times\mathcal{Y}}$ to $\mathcal{X}$ or $\mathcal{Y}$. The projection to $\mathcal{X}$ is defined as:

$$m^{\mathcal{X}\times\mathcal{Y}\downarrow\mathcal{X}}(A) = \sum_{B\downarrow A} m^{\mathcal{X}\times\mathcal{Y}}(B), \tag{30}$$

where the summation is carried out over all $B \subseteq \mathcal{X} \times \mathcal{Y}$ such that by elimination of $\mathcal{Y}$, B reduces to $A \subseteq \mathcal{X}$.

Testing scenario 6. Consider two spaces: (i) weather $\mathcal{X} = \{g, \bar{g}\}$, where g stands for good and $\bar{g}$ for the opposite; (ii) location $\mathcal{Y} = \{b, h, o\}$, where b, h and o stand for beach, home and office, respectively. Prior knowledge comes in the form of the implication rule: if the weather is good, the subject is located at the beach, with confidence 0.8. Weather forecast for tomorrow is *good* with the confidence 0.7. Question: what is the probability that the subject will be at the beach tomorrow?■

The product space has six elements, that is

$$\mathcal{X} \times \mathcal{Y} = \{(g,b),(\bar{g},b),(g,h),(\bar{g},h),(g,o),(\bar{g},o)\}.$$

The TBM expresses belief over the power set $2^{\mathcal{X}\times\mathcal{Y}}$. In order to combine in TBM the two pieces of information (the weather forecast and the implication rule), both must be expressed by bbas over the product space. The resulting bbas contain only a limited number of focal sets, and are shown in columns $m_1^{\mathcal{X}\times\mathcal{Y}}$ and $m_2^{\mathcal{X}\times\mathcal{Y}}$ of Table 13. The bba resulting from the conjunctive combination (20), is given in the last column of Table 13. Finally, in order to answer the question posed in the Testing scenario 6, we marginalise $m_{12}^{\mathcal{X}\times\mathcal{Y}}$ to $\mathcal{Y}$ and obtain that the resulting bba $m_{12}^{\mathcal{X}\times\mathcal{Y}\downarrow\mathcal{Y}}$ has two focal sets: $\{b\}$ and $\mathcal{Y}$, with assigned basic belief masses of 0.56 and 0.44, respectively. After applying the pignistic transform (24) to $m_{12}^{\mathcal{X}\times\mathcal{Y}\downarrow\mathcal{Y}}$ we find that the subject will be on the beach tomorrow with pignistic probability 0.71. However, Voorbraak probability of the same event is only 0.53.

focal sets	weather forecast $m_1^{\mathcal{X}\times\mathcal{Y}}$	implication $m_2^{\mathcal{X}\times\mathcal{Y}}$	Combination $m_{12}^{\mathcal{X}\times\mathcal{Y}}$
$\{(g,b)\}$			0.56
$\{(g,b),(g,h),(g,o)\}$	0.7		0.14
$\{(g,b),(\bar{g},b),(\bar{g},h),(\bar{g},o)\}$		0.8	0.24
$\mathcal{X}\times\mathcal{Y}$	0.3	0.2	0.06

Table 13: Testing scenario 6 - TBM

There is no clear guideline as to how to apply the probabilistic or the possibilistic approach to the Testing scenario 6. One possibilistic solution is given in Table 14. In order to answer the posed question, we need to project $\pi_1 \cdot \pi_2$, given in the 3rd column, to the subspace $\mathcal{Y}$. The resulting possibility function over $\mathcal{Y}$ is given by $\pi(b) = \max(1, 03/0.7) = 1$, $\pi(h) = \pi(o) = \max(1/4, 0.3/0.7) = 0.3/0.7$. Using transformation (12), the probability that the subject will be on the beach tomorrow is 0.54.

	weather forecast	implication	combination
	π_1	π_2	$\pi_1 \cdot \pi_2$
(g,b)	1	1	1
(g,h)	1	$(1-0.8)/0.8$	$1/4$
(g,o)	1	$(1-0.8)/0.8$	$1/4$
$(\bar{g},b)$	$0.3/0.7$	1	$0.3/0.7$
$(\bar{g},h)$	$0.3/0.7$	1	$0.3/0.7$
$(\bar{g},o)$	$0.3/0.7$	1	$0.3/0.7$

Table 14: Testing scenario 6 - possibilistic approach

4.3. Dezert-Smarandache approach

There are many developments and interpretations of the belief function theory. One noteworthy extension is the Dezert-Smarandache theory (DSmT) [42]. The main feature of the DSmT is that it relaxes the assumption that the elements of the space of classes $\mathcal{X}$ are mutually exclusive. There are situations where this generalisation is indeed relevant. For example, constructing a discrete space $\mathcal{X}$ using fuzzy and relative concepts over continuous spaces, such as size (small/large) or danger (low/medium/high), could lead to non-exclusive elements. Another example from practice is the full set of NATO affiliations (allegiances) which, for example, contains as elements both the class *assumed friend* and the class *friend*. Exclusiveness of such a set is questionable: *assumed friend* is also a *friend*, with a difference being the degree of confidence. Numerous other practical examples of non-excluded classes can be found in [43].

In the DSmT, the belief is expressed over the hyper-power set, denoted $D^{\mathcal{X}}$. The hyper-power set consists of all compositions built from the elements of $\mathcal{X}$ using union and intersection operations. For example, if $N = 2$ and $\mathcal{X} = \{x_1, x_2\}$, then $D^{\mathcal{X}} = \{\alpha_i\}_{0 \leq i \leq 4}$, where $\alpha_0 = \emptyset$, $\alpha_1 = x_1$, $\alpha_2 = x_2$, $\alpha_3 = x_1 \cup x_2$ and $\alpha_4 = x_1 \cap x_2$. For $N = 3$, the hyper-power set becomes significantly more complicated, because it includes the elements such as $(x_1 \cup x_2) \cap x_3$, $(x_1 \cap x_2) \cup x_3$, and likewise. The cardinality of $D^{\mathcal{X}}$ grows rapidly with N, as 2^{2^N}. Thus for $N = 1, 2, 3, 4, 5, 6$, we have $|D^{\mathcal{X}}| = 1, 2, 5, 19, 167, 7580$, respectively. Note that, in general, a complement of an element of $D^{\mathcal{X}}$ is not included in $D^{\mathcal{X}}$.

The central element of the DSmT is the generalised bba $\mu : D^{\mathcal{X}} \to [0, 1]$, characterised by $\mu(\emptyset) = 0$ and $\sum_{A \in D^{\mathcal{X}}} \mu(A) = 1$. One can now introduce the conjunctive combination (referred to as the classic DSmT rule of combination [42]) of two generalised bbas μ_1 and μ_2, assuming they are defined over the same space of classes $\mathcal{X}$, and coming from two independent sources. The conjunctive combination, expressed as $\mu_{12} = \mu_1 \bigoplus \mu_2$, is defined as:

$$\mu_{12}(A) = \sum_{\substack{B,C \in D^{\mathcal{X}} \\ B \cap C = A}} \mu_1(B) \cdot \mu_2(C), \qquad \forall A \in D^{\mathcal{X}}. \tag{31}$$

The output of the combination rule (31) is guaranteed to be another generalised bba, that is $\mu_{12}(\emptyset) = 0$ and $\sum_{A \in D^{\mathcal{X}}} \mu_{12}(A) = 1$.

If one is confident that the elements of the space of classes $\mathcal{X}$ are mutually exclusive, the hyper-power set reduces to the power set, and (31) reduces to conjunctive combination (20). There is also a hybrid model in DSmT, where some elements in $\mathcal{X}$ are treated as mutually exclusive, while others are not. A great deal of the DSmT is devoted to hybrid models and new combination rules, such as the *partial conflict redistribution* rules [44].

5. Imprecise probabilities

5.1. Fundamentals

Imprecise (or *indeterminate*) probabilities provide a general framework for modelling uncertain knowledge. Despite being termed imprecise, this approach aims to model uncertainty in a more precise manner by introducing upper and lower probabilities or more generally upper and lower previsions [24]. Informally, to paraphrase Coolen *et al* [45], the lower probability of an event A can be interpreted as reflecting the certain evidence in favour of A, whilst the upper probability reflects all the evidence possibly in favour of A. The difference between these probabilities thus reflects the imprecision, or lack of perfect probabilistic information, relating to A. Accordingly, belief functions [29, 30], random sets [8] and possibility measures [19, 20, 22] can also be considered as special cases of imprecise probabilities. The advantage of this more general framework is the ability of imprecise probability models to allow for indecision in order to avoid making an incorrect decision.

The main theoretical grounding underpinning imprecise probabilities stems from Walley's [24] theory of *coherent lower previsions*[13]. His work approaches subjective probability and imprecision from a behavioural point of view resulting in three central concepts: avoiding sure-loss, coherence and natural extension.

[13] It is worth noting that Walley's theories were influenced by de Finetti's [46] work on imprecise subjective probability and Williams [47] work on conditional previsions

Using this theoretical framework, imprecise probability models have found application in areas such as classification [48], information fusion [49], engineering analysis [50] and tracking [51]. In the following, we briefly summarize Walley's work; however, for a detailed explanation we refer the reader to [24, 52, 53].

Referring to Sec. 2.1, let a discrete variable X be defined over $\mathcal{X}$. Suppose that the evidence about X can not be represented by a single PMF p, but rather by a closed-convex set $K(X)$ of probability functions, referred to as the *credal set.*

Consider a (bounded) function $f : \mathcal{X} \to \mathbb{R}$, such that $f \in \mathcal{L}(\mathcal{X})$, where $\mathcal{L}(\mathcal{X})$ denotes the linear space of all functions on $\mathcal{X}$ [24, Ch.1]. For a single probability function $p \in K(X)$, the expectation of f is defined as:

$$E[f] = \sum_{x_i \in \mathcal{X}} f(x_i)p(x_i). \tag{32}$$

Using this definition, for each $p \in K(X)$ we have an associated $E[f]$, which, due to the linearity of the expectation operator, satisfies the following equation: $\underline{E}[f] \leq E[f] \leq \overline{E}[f]$, where $\underline{E}[f]$ is the coherent lower prevision (CLP), or lower expectation of f, defined as:

$$\underline{E}[f] = \min_{p \in K(X)} \sum_{x_i \in \mathcal{X}} f(x_i)p(x_i), \quad \forall f \in \mathcal{L}(\mathcal{X}). \tag{33}$$

Similarly, $\overline{E}[f] = \max_{p \in K(X)} \sum_{x_i \in \mathcal{X}} f(x_i)p(x_i) = -\underline{E}[-f]$, is the upper expectation. There are two extreme cases: (i) when credal set $K(X)$ includes only one PMF p and (ii) when $K(X)$ includes all possible PMFs. The first case corresponds to the most informative situation for which $\underline{E}[f] = E[f] = \overline{E}[f]$. The second case, referred to as *vacuous*, results in the CLP $\underline{E}[f] = \min_{x_i \in \mathcal{X}} f(x_i)$.

The CLP (33) can represent the lower probability of an event $A \subseteq \mathcal{X}$ by setting f to be the indicator function over A, i.e. $f = I_A$. Then from (33) we have:

$$\begin{aligned} \underline{E}[I_A] &= \min_{p \in K(X)} \sum_{x_i \in \mathcal{X}} I_A(x_i)p(x_i) && (34) \\ &= \min_{p \in K(X)} \sum_{x_i \in A} p(x_i) = \underline{P}(A) && (35) \end{aligned}$$

where $\underline{P}(A)$ is the lower probability of A. Similarly, $\overline{E}[I_A] = \overline{P}(A)$ is the upper probability of A. Because $P(A) = 1 - P(A^c)$, we have that:

$$\overline{P}(A) = \max_{p \in K(X)} \left(1 - \sum_{x_i \in A^c} p(x_i)\right) = 1 - \min_{p \in K(X)} \sum_{x_i \in A^c} p(x_i) = 1 - \underline{P}(A^c). \tag{36}$$

Note that if $K(X)$ includes only one probability function (i.e. the probability is precise), then $\underline{P}(A) = P(A) = \overline{P}(A)$, and we obtain the classical probabilistic framework. The other extreme, when the credal set includes all possible PMFs, results in $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$.

In the context of classification, we need to introduce the feature variable Z taking values in $\mathcal{Z}$. The *conditional lower prevision* for a function $h \in \mathcal{L}(\mathcal{X} \times \mathcal{Z})$ is denoted $\underline{E}[h|x_i]$; it defines the lower expectation of h w.r.t Z conditioned on $X = x_i$. This conditional lower prevision is calculated simply by replacing $p \in K(X)$ with $g(\cdot|x) \in K(Z|x)$, where $K(Z|x)$ is the convex set of likelihood functions (conditional probability functions).

Relation to belief functions. The lower probability of event $A \in \mathcal{X}$, i.e. $\underline{P}(A)$, was introduced in (35). Recall also that *bel* and *pl* were introduced in (17) and (19), respectively. If $\underline{P}$ satisfies these properties: (i) $\underline{P}(\emptyset) = 0$; (ii) $\underline{P}(\mathcal{X}) = 1$; (iii) ∞-monotonicity, see (18), then the lower probability $\underline{P}$ is a belief function *bel* and its conjugate upper probability $\overline{P}$ is a plausibility function *pl*. Belief functions are thus a special case of CLPs. The credal set $K(X)$ associated with a bba m is the set of all probability functions consistent with m; it can be generated by a procedure described in [54]. According to the behavioural interpretation of lower and upper expectations (in terms of buying and selling prices on gambles), Walley [24, Ch.5] showed that the normalised conjunctive rule, i.e. (20) with (21), is incompatible with the coherence approach, because it can incur a sure loss.

5.2. The analog of Bayes rule

The analog of (2) in the framework of imprecise probabilities is formulated as follows. Given a prior probability function p over $\mathcal{X}$ and a feature measurement $z \in \mathcal{Z}$, the posterior lower expectation of f is computed as:

$$\underline{E}[f|z] = \min_{\substack{p \in K(X); g \in K(Z|x) \\ \text{s.t. } \sum_{x \in \mathcal{X}} g(z|x)p(x) > 0}} \frac{\sum_{x \in \mathcal{X}} f(x)g(z|x)p(x)}{\sum_{x \in \mathcal{X}} g(z|x)p(x)}, \tag{37}$$

where we have assumed that there exists at least one $p \in K(X)$ and one $g \in K(Z|x)$ such that $\sum_{x \in \mathcal{X}} g(z|x)p(x) > 0$ (if this is not the case, the likelihood and prior beliefs are in conflict). Note that in the formulation (37), both the likelihood function $g(\cdot|x)$ and the prior p are treated as being imprecise. If, however, the prior is precise, then minimisation in (37) is carried out only w.r.t. $g(\cdot|x) \in K(Z|x)$. Likewise, if the likelihood is precise, minimisation is carried out only w.r.t. $p \in K(X)$.

Let us reformulate optimisation (37) by introducing a (scalar) variable ν, as follows:

$$\begin{aligned}\frac{\sum_{x\in\mathcal{X}} f(x)g(z|x)p(x)}{\sum_{x\in\mathcal{X}} g(z|x)p(x)} = \nu \quad &\Rightarrow \quad \sum_{x\in\mathcal{X}} f(x)g(z|x)p(x) = \nu \sum_{x\in\mathcal{X}} g(z|x)p(x) \\ &\Rightarrow \quad \sum_{x\in\mathcal{X}} (f(x)-\nu)g(z|x)p(x) = 0.\end{aligned}$$

The problem (37) can then be expressed as [24, Appendix J.1]:

$$\underline{E}[f|z] = \max \nu, \quad s.t. \quad \min_{\substack{g\in K(Z|x)\\ p\in K(X)}} \sum_{x\in\mathcal{X}} (f(x)-\nu)g(z|x)p(x) \geq 0. \tag{38}$$

Note that (38) involves two optimisation problems, a minimisation and a maximisation. Given ν, the minimisation problem is *bilinear* in the unknown $g(\cdot|x)$ and p. If for example p is precise, then the minimisation problem becomes linear in the unknown $g(\cdot|x)$ and can efficiently be solved by linear programming. Bilinear optimisation, on the other hand, can be solved using for example the approach discussed in [55]. Maximisation over ν in (38) can be solved by a bisection method.

Decision making. The question is how to make a decision using the interval valued expectations created by imprecise probability models? There are several decision criteria that can be used, such as *maximality*, *E-admissibility*, *interval dominance*, Γ-*maximax*, Γ-*maximin* [24, Ch.3], see [56] for a recent review. E-admissibility, maximality, and interval dominance have the nice property that the more determinate our beliefs (i.e., the smaller is the credal set), the smaller the set of optimal decisions. Conversely, Γ-*maximax* and Γ-*maximin* lack this property, and usually only select a single decision, even in the case of complete ignorance. Interval dominance is easy to compute but much weaker than E-admissibility and maximality. E-admissibility and maximality are very similar, but with subtle differences. The decision criterion, for a particular application, depends on the goals of the decision maker (what properties should optimality satisfy?), and possibly also on the size and structure of the problem if computational issues arise [56]. In this paper, we have selected maximality because it is slightly weaker than E-admissibility and easier to understand. Under this criterion, an action a_i, which defines an utility function f_{a_i}, dominates (or is preferred to) another action a_j, which defines an utility function f_{a_j}, if for all $p \in K(X)$:

$$E^p(f_{a_i} - f_{a_j}) > 0, \tag{39}$$

where E^p denotes the expectation w.r.t the probability p. A necessary and sufficient condition for (39) to be satisfied is that [48]

$$\underline{E}(f_{a_i} - f_{a_j}) > 0. \tag{40}$$

Maximality criterion can be extended to conditional and posterior expectations.

Testing scenario 3. The framework of imprecise probabilities (IP) allows us to naturally map the beliefs (a)-(c) of the Testing scenario 3 into three conditional credal sets:
(a) $K(Z|x_1) = \{g(\cdot|x_1) : g(\zeta_1|x_1) = 1, g(\zeta_2|x_1) = 0, g(\zeta_3|x_1) = 0\}$;
(b) $K(Z|x_2) = \{g(\cdot|x_2) : g(\zeta_1|x_2) + g(\zeta_2|x_2) = 1,\ g(\zeta_3|x_2) = 0\}$;
(c) $K(Z|x_3) = \{g(\cdot|x_3) : g(\zeta_1|x_3) + g(\zeta_2|x_3) + g(\zeta_3|x_3) = 1\}$.
Observe that, $K(Z|x_1)$ includes a single PMF (it means that, given x_1, we are sure that the measurement is ζ_1), $K(Z|x_2)$ includes all conditional PMFs $g(\cdot|x_2)$ such that $g(\zeta_3|x_2) = 0$ (i.e. given x_2, we only know that ζ_3 is impossible); $K(Z|x_3)$ includes all possible conditional PMFs $g(\cdot|x_3)$ (i.e. given x_3, any feature measurement is possible).

Testing scenario 3 involves application of (38) sequentially, for $k = 1, 2, \ldots, 17$. By choosing $f = I_{\{x_1\}}$, we compute the posterior lower probability of class x_1, while, according to (36), for $f = 1 - I_{\{x_2,x_3\}}$, we compute the posterior upper probability of class x_1. The case $k = 1$ is different and simpler than $k = 2, 3, \ldots, 17$, because the prior PMF p is precise. Recall that at $k = 1$, the feature measurement is $z_1 = \zeta_1$. Then one can write the minimisation problem in (38) for $f(x) = I_{\{x_1\}}(x)$ as the following linear program:

$$\begin{array}{l}
\min_g\ (1-\nu)g(\zeta_1|x_1)p(x_1) + (-\nu)g(\zeta_1|x_2)p(x_2) + (-\nu)g(\zeta_1|x_3)p(x_3)\\
\text{subject to}\\
g(\zeta_1|x_1) = 1,\ g(\zeta_2|x_1) = 0,\ g(\zeta_3|x_1) = 0,\\
g(\zeta_1|x_2) + g(\zeta_2|x_2) = 1,\ g(\zeta_3|x_2) = 0,\\
g(\zeta_1|x_3) + g(\zeta_2|x_3) + g(\zeta_3|x_3) = 1,\\
g(\zeta_i|x_j) \geq 0,\ \text{for } i, j = 1, 2, 3.
\end{array} \tag{41}$$

where ν is given. In matrix form, (41) can be written as:

$$\min_{\mathbf{g}}\ \mathbf{c}^{\mathsf{T}}\mathbf{g} \text{ subject to } \mathbf{A}\mathbf{g} \leq \mathbf{b} \text{ and } \mathbf{g} \geq \mathbf{0} \tag{42}$$

where

$$\mathbf{c} \; = \; \begin{bmatrix} (1-\nu)p(x_1) \\ 0 \\ 0 \\ -\nu p(x_2) \\ 0 \\ 0 \\ -\nu p(x_3) \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{g} = \begin{bmatrix} g(\zeta_1|x_1) \\ g(\zeta_2|x_1) \\ g(\zeta_3|x_1) \\ g(\zeta_1|x_2) \\ g(\zeta_2|x_2) \\ g(\zeta_3|x_2) \\ g(\zeta_1|x_3) \\ g(\zeta_2|x_3) \\ g(\zeta_3|x_3) \end{bmatrix}$$

$$\mathbf{A} \; = \; \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ -1 \\ 0 \\ -1 \end{bmatrix} \tag{43}$$

In order to maximise ν, see (38), in each iteration of the bisection method one needs to solve the described linear program.

For illustration, let us derive the lower/upper posterior probability of class x_1 at $k = 1$. Taking into account the simplification described above and using $z_1 = \zeta_1$, we have from (38):

$$\underline{p}(x_1|\zeta_1) = \max \nu \text{ s.t. } \min \frac{1}{3}[(1-\nu)g(\zeta_1|x_1) - \nu\, g(\zeta_1|x_2) - \nu\, g(\zeta_1|x_3)] \geq 0 \tag{44}$$

Note that $(1-\nu)g(\zeta_1|x_1)-\nu\, g(\zeta_1|x_2)-\nu\, g(\zeta_1|x_3)$ has a minimum when $g(\zeta_1|x_2) = 1$ and $g(\zeta_1|x_3) = 1$, that is, the condition in (44) is given by:

$$(1-\nu)g(\zeta_1|x_1) - \nu\, g(\zeta_1|x_2) - \nu\, g(\zeta_1|x_3) \geq 1 - 3\nu \geq 0.$$

Therefore $\nu \leq 1/3$, that is the maximum value of ν is $1/3$. Thus the lower posterior probability of class x_1 is $\underline{p}(x_1|\zeta_1) = 1/3$.

In order to derive the upper posterior probability of class x_1, according to

(36), first we must determine:

$$\underline{p}(\{x_2, x_3\}|\zeta_1) = \max \nu \text{ s.t. } \min \frac{1}{3}[-\nu\, g(\zeta_1|x_1) + (1-\nu) g(\zeta_1|x_2) + (1-\nu) g(\zeta_1|x_3)] \geq 0 \quad (45)$$

The condition in (45) has the minimum when $g(\zeta_1|x_2) = 0$ and $g(\zeta_1|x_3) = 0$, that is:

$$-\nu\, g(\zeta_1|x_1) + (1-\nu) g(\zeta_1|x_2) + (1-\nu) g(\zeta_1|x_3) \geq -\nu \geq 0.$$

Hence the maximum value of ν is 0, i.e. $\underline{p}(\{x_2, x_3\}|\zeta_1) = 0$. From (36) it follows that $\overline{p}(x_1|\zeta_1) = 1 - \underline{p}(\{x_2, x_3\}|\zeta_1) = 1$.

At $k = 2, 3, \dots$, the prior is replaced with the posterior from the previous time $k-1$, that is, with the IP $p(\cdot|z_{1:k-1})$. Thus, for $k = 2, 3, \dots$, (38) takes the following form:

$$\underline{p}(x_i|z_{1:k}) = \max \nu, \quad s.t. \quad \min_{\substack{g \in K(Z|x) \\ p \in K_{k-1}(X)}} \sum_{x \in \mathcal{X}} (I_{\{x_i\}}(x) - \nu) g(z_k|x) p(x|z_{1:k-1}) \geq 0. \quad (46)$$

where $K_{k-1}(X)$ is the credal set of the imprecise posterior PMF $p(\cdot|z_{1:k-1})$. This bilinear minimisation problem can be solved, for example, by finding the vertices of the credal set $K_{k-1}(X)$ by means of random search directions. In doing so, the posterior lower and upper probabilities, $\underline{p}(\cdot|z_{1:k-1})$ and $\overline{p}(\cdot|z_{1:k-1})$, respectively, are used as the bounds for a new linear program. By solving this linear program with random search directions, we determine the vertices of $K_{k-1}(X)$. Details are given in the MATLAB routine script_14, which implements the IP solution for Testing scenario 3.

The lower/upper (LP/UP) posterior probabilities, computed using the sequence of 17 feature measurements available for classification (recall that all measurements in this sequence are ζ_1, except $z_7 = \zeta_2$ and $z_{13} = \zeta_3$), are shown in Fig. 10. The following observations and subsequent decisions using the maximality criterion, can be made (we also give the intuition behind the decisions but we refer the reader to the definition of maximality and the previous derivations for really understanding what is going on).

- From $k = 1$ to $k = 6$ the LP of class x_1 is 1/3, while the UP is 1. Since the UPs of classes x_2 and x_3 exceed the LP of class x_1, and vice versa, none of the classes dominates the others. Hence, during this interval, the classification decision is the entire space $\mathcal{X}$ (effectively, indecision).

- From time $k = 7$, the UP of class x_1 is zero, and therefore, both class x_2 and class x_3 dominate class x_1. Hence, in the interval from $k = 7$ to $k = 12$, the classification decision is the set $\{x_2, x_3\}$.

- From $k = 13$, the UP of class x_2 is zero too, and therefore the class decision is x_3.

As discussed previously, these are the right decisions in this scenario.

MATLAB exercise. $\gg$ script_14;

Testing scenario 4 examines the classification performance in the case where the confusion matrix is mismatched. When using imprecise probabilities, however, we are not restricted to precise, possibly mismatched, likelihoods. Instead, the likelihoods can be specified as the confidence intervals, thereby (potentially) avoiding the model mismatch situations. This is illustrated with the next example.

Modified testing scenario 4. The confusion matrix used in the generation of feature measurements is the same as in Table 1.(a). A sequence of $K = 20$ feature measurements is generated at random using this confusion matrix, with the *true* class being x_2. The confusion matrix available for classification is imprecise, and given by Table 15. Note that the correct values, which appear in Table 1.(a), are contained in the intervals of Table 15 (i.e. Table 15 satisfies modeling integrity). The IP classifier for this testing scenario can be built following the same method presented earlier in solving Testing scenario 3. Assuming the prior class probabilities are precise, i.e. $p(x_1) = p(x_2) = 2/5$ and $p(x_3) = 1/5$, we want to determine the average performance of the IP classifier, obtained from 100 independent Monte Carlo runs. Fig. 11 shows the classification results. The red solid and dashed lines indicate the average lower and upper posterior probabilities of class x_2, respectively (x_2 is the true class). The blue line is shown only for verification, because it represents the output of the (standard) Bayes classifier which *knows* the true confusion matrix given by Table 1.(a). The IP approach provides reliable classification: the LP/UP interval always includes

Table 15: The imprecise confusion matrix available for classification

$g(\zeta_i \vert x_j)$	x_1	x_2	x_3
ζ_1	$[0.65, 0.75]$	$[0.10, 0.20]$	$[0.10, 0.20]$
ζ_2	$[0.10, 0.20]$	$[0.65, 0.75]$	$[0.10, 0.20]$
ζ_3	$[0.10, 0.20]$	$[0.10, 0.20]$	$[0.65, 0.75]$

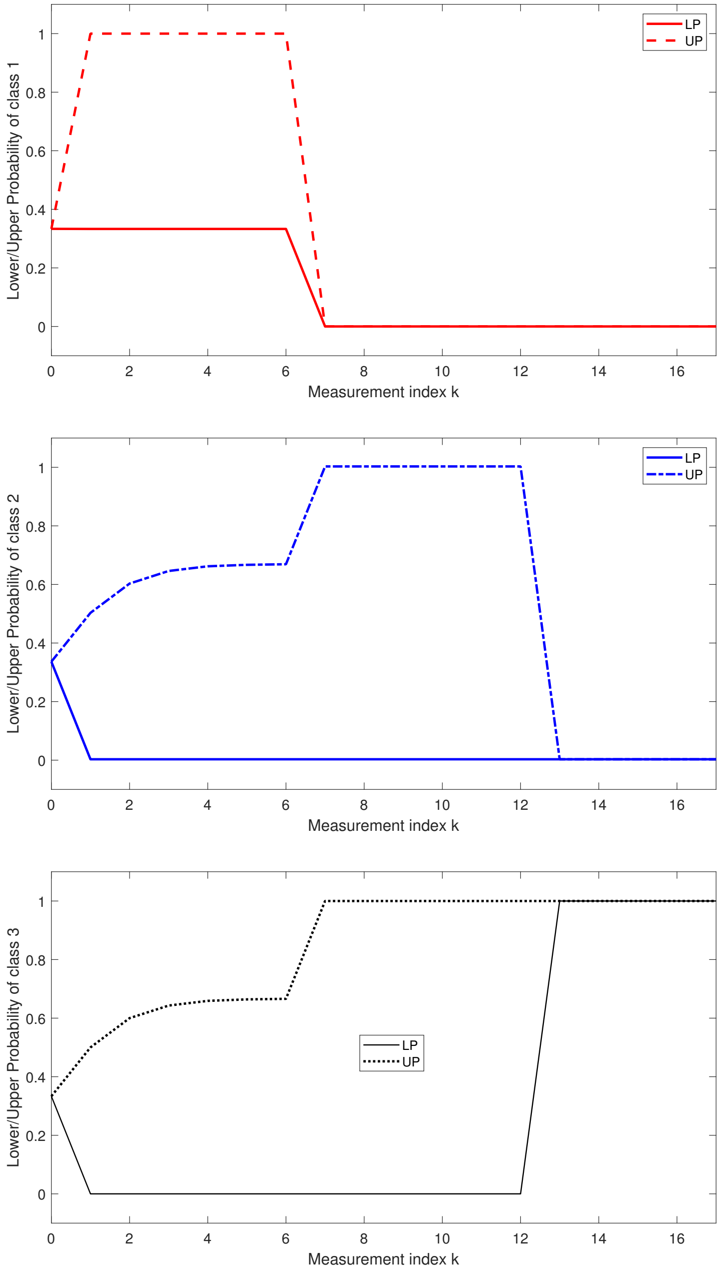

Figure 10: The lower/upper posterior class probabilities versus k for the Testing scenario 3

the blue line (achieved in ideal circumstances). Observe also that the IP classifier is cautious, in the sense that there is a gap between the lower and upper probabilities. This gap could be reduced by making the intervals in Table 15 tighter (less uncertain).

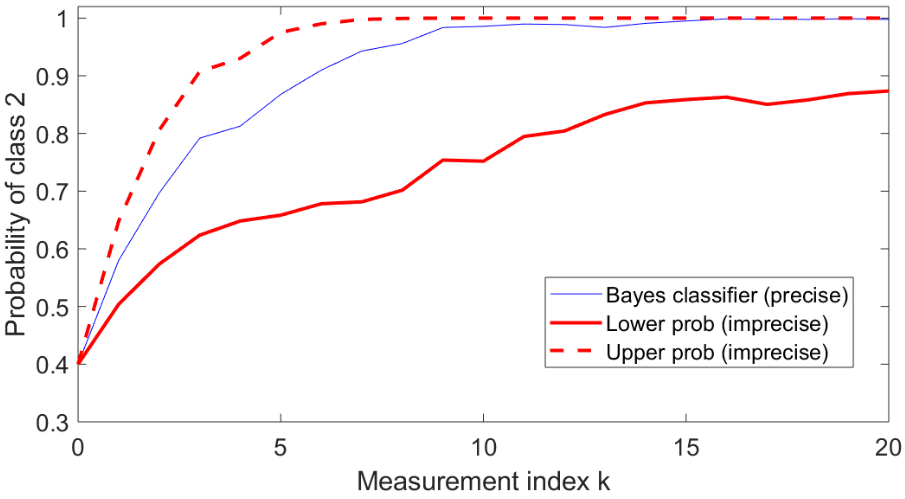

Figure 11: Classification with the imprecise confusion matrix in the modified Testing scenario 4. The red lines represent the lower/upper posterior probabilities of class x_2 (true class) versus k, averaged over 100 Monte Carlo runs. The blue line is output of the Bayesian classifier which *knows* the true confusion matrix.

MATLAB exercise. $\gg$ script_15(100);

Multi-source combination. Imprecise probabilities provide a flexible framework for modelling and aggregating the outputs of multiple classifiers or the opinions of several experts. First, they allow for a more reliable representation of expert information (the experts are not forced to specify a single probability measure in order to represent their knowledge). Second, they provide a natural setting for modelling conflicting opinions, using imprecision as a means of expressing disagreement between different opinions. The analog of (4) is referred to as the *conjunction rule*. Suppose two classifiers (or experts) express their opinion about X in the form of credal sets $K_1(X)$ and $K_2(X)$, respectively. The conjunction rule is defined as the convex hull of the intersection of $K_1(X)$ and $K_2(X)$. In this way, conjunction aims at gaining as much information as possible from each of the experts: the aggregate (the output of the rule) is at least as informative as each of the experts' credal set. The conjunction, however, may not exist. In particular, when the two classifiers (experts) make conflicting statements, the resulting intersection is an empty set. In this case, various alternatives exists, one of them being the *unanimity rule*, defined as the convex hull of the union of the credal sets. Both rules can be easily implemented in software. For example, in the case of the conjunction rule, we simply impose all the constraints defining credal sets $K_1(X)$ and $K_2(X)$ into one linear program. For more details we refer the reader to [57, 58, 59].

6. Concluding remarks

In solving practical problems involving imperfect (uncertain) domain knowledge and/or input data, it is important to use the mathematical models characterised by *integrity*, meaning that the actual knowledge is accurately represented (without additional assumptions). Imperfect information can be a consequence of stochastic variability, imprecision, or a model-mismatch.

This tutorial paper reviewed some of the prevalent approaches to quantitative modeling of uncertain information and reasoning for model-based classification. They included the Bayesian approach (standard and Mahler's), the approaches based on possibility theory, the belief function theory and the imprecise probability theory. A rough guideline on the computation time (in seconds) is provided in Table 16. The computation time was estimated using Testing scenario 3.

Table 16: Estimated computation time (in seconds) for Testing scenario 3

standard Bayesian (script_3.m)	Mahler Bayesian (script_4.m)	Possibility functions (script_7.m)	Belief functions (script_11.m)	Imprecise probabilities (script_14.m)
0.2	0.2	0.2	0.4	83.7

There is no consensus at present on the best approach. Quantitative reasoning under uncertainty is an active research field and may converge to a unified theory in the future. The imprecise probability theory provides a rich enough framework to lead to a unified theory[14], however, the increase in computation is significant due to numerical optimisation.

[14] Flexible software tools for inference over large credal networks are available, see [60]

Acknowledgments

This research was partly supported by DST Group under the Research Agreement "Classification decisions under uncertainty".

References

[1] A. Motro. Imprecision and uncertainty in database systems. In *Fuzziness in Database Management Systems*, pages 3–22, Heidelberg, 1995. Physica-Verlag HD.

[2] Y. Li, J. Chen, and L. Feng. Dealing with uncertainty: A survey of theories and practices. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2463–2482, 2013.

[3] P. Smets. Imperfect information: Imprecision and uncertainty. In *Uncertainty management in information systems*, pages 225–254. Springer, 1997.

[4] R. Kruse, E. Schwecke, and J. Heinsohn. *Uncertainty and vagueness in knowledge based systems: numerical methods.* Springer, 2012.

[5] D. M Buede and P. Girardi. A target identification comparison of Bayesian and Dempster-Shafer multisensor fusion. *IEEE Trans on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 27(5):569–577, 1997.

[6] S. Maskell. A Bayesian approach to fusing uncertain, imprecise and conflicting information. *Information Fusion*, 9(2):259–277, 2008.

[7] T. Denoeux. Introduction to belief functions. https://www.hds.utc.fr/~tdenoeux/dokuwiki/_media/en/lecture1.pdf, 2017. Accessed: 8th Nov. 2018.

[8] R. Mahler. *Statistical Multisource Multitarget Information Fusion.* Artech House, 2007.

[9] C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

[10] D. Barber. *Bayesian reasoning and machine learning.* Cambridge University Press, 2012.

[11] J. Llinas and C.-Y. Chong. Object classification in a distributed environment. In J. Llinas M. Liggins II D. Hall, C.-Y. Chong, editor, *Distributed data fusion for network-centric operations*, chapter 9. CRC Press, 2017.

[12] L. Zadeh. On the validity of Dempster's rule of combination of evidence. Memo M 79/24, 1979.

[13] R. P. S. Mahler. Optimal/robust distributed data fusion: a unified approach. In *Proc. SPIE: Signal Processing, Sensor Fusion, and Target Recognition IX*, volume 4052, pages 128–139, 2000.

[14] M. B. Hurley. An information theoretic justification for covariance intersection and its generalization. In *Proc. of the 5th Intern. Conf. Information Fusion*, volume 1, pages 505–511. IEEE, 2002.

[15] D. J. C. MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

[16] B. Ristic and P. Smets. Target identification using belief functions and implication rules. *IEEE transactions on Aerospace and Electronic Systems*, 41(3):1097–1103, 2005.

[17] B. Ristic. Target classification with imprecise likelihoods: Mahler's approach. *IEEE Trans. on Aerospace and Electronic Systems*, 47(2):1530–1534, 2011.

[18] R. Mahler and A. El-Fallah. The random set approach to nontraditional measurements is rigorously Bayesian. In *Proc. SPIE: Signal Processing, Sensor Fusion, and Target Recognition XXI*, volume 8392, page 83920D, 2012.

[19] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[20] D. Dubois and H. Prade. *Possibility theory: An approach to computerised processing of uncertainty.* Plenum Pub., 1988.

[21] D. Dubois and H. Prade. Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of mathematics and Artificial Intelligence*, 32(1-4):35–66, 2001.

[22] D. Dubois and H. Prade. Possibility theory and its applications: Where do we stand? In *Springer Handbook of Computational Intelligence*, pages 31–60. Springer, 2015.

[23] F. Hampel. Nonadditive probabilities in statistics. *Journal of Statistical Theory and Practice*, 3(1):11–23, 2009.

[24] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman & Hall, 1991.

[25] M. Boughanem, A. Brini, and D. Dubois. Possibilistic networks for information retrieval. *Intern. Journal of Approximate Reasoning*, 50(7):957–968, 2009.

[26] J. Houssineau and A. N. Bishop. Smoothing and filtering with a class of outer measures. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):845–866, 2018.

[27] P. P. Shenoy. Using possibility theory in expert systems. *Fuzzy Sets and Systems*, 52(2):129–142, 1992.

[28] B. Ristic, J. Houssineau, and S. Arulampalam. Robust target motion analysis using the possibility particle filter. *IET Radar Sonar Navigation*, 2018. (in print).

[29] G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[30] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, pages 325–339, 1967.

[31] P. Smets and R. Kennes. The transferable belief model. *Artificial intelligence*, 66(2):191–234, 1994.

[32] P. Smets. The transferable belief model for quantified belief representation. In *Quantified Representation of Uncertainty and Imprecision*, pages 267–301. Springer, 1998.

[33] F. Delmotte and P. Smets. Target identification based on the transferable belief model interpretation of Dempster-Shafer model. *IEEE Trans. on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 34(4):457–471, 2004.

[34] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 633–664. Springer, 2008.

[35] P. L. Bogler. Shafer-Dempster reasoning with applications to multisensor target identification systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(6):968–977, 1987.

[36] B. Ristic and P. Smets. The TBM global distance measure for the association of uncertain combat ID declarations. *Information fusion*, 7(3):276–284, 2006.

[37] B. Ristic, M. C. Florea, and É. Bossé. Addendum for "the TBM global distance measure for the association of uncertain combat ID declarations". *Information Fusion*, 18:197–198, 2014.

[38] T. Denoeux. Decision-making with belief functions: A review. *International Journal of Approximate Reasoning*, 109:87–110, 2019.

[39] P. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *Intern. Journal of Approximate Reasoning*, 38(2):133–147, 2005.

[40] Frans Voorbraak. A computationally efficient approximation of dempster-shafer theory. *Logic Group Preprint Series*, 35, 1988.

[41] B. Ristic and P. Smets. Fusion of uncertain combat identity declarations and implication rules using the belief function theory. In *Proc. SPIE, Signal and Data Processing of Small Targets 2005*, volume 5913, page 591318, 2005.

[42] F Smarandache and J Dezert. *Advances and applications of DSmT for information fusion.* ARP, 2004.

[43] F. Smarandache and J. Dezert. *Advances and Applications of DSmT for Information Fusion, Vol. IV: Collected Works.* Infinite Study, 2015.

[44] F. Smarandache and J. Dezert. Information fusion based on new proportional conflict redistribution rules. In *8th Intern. Conf. on Information Fusion*, volume 2, 2005.

[45] F. P. A. Coolen, M. C. M. Troffaes, and T. Augustin. Imprecise probability. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 645–648. Springer, 2011.

[46] B. de Finetti. *Theory of Probability.* Wiley, London, 1974.

[47] P. Williams. Notes on conditional previsions. *Intern. Journal of Approximate Reasoning*, 44(3):366 – 383, 2007. Initially an unpublished technical report written in 1975.

[48] A. Benavoli and B. Ristic. Classification with imprecise likelihoods: A comparison of TBM, random set and imprecise probability approach. In *Proc. of the 14th Intern. Conf. on Information Fusion,*, pages 1–8. IEEE, 2011.

[49] A. Benavoli, M. Zaffalon, and E. Miranda. Robust filtering through coherent lower previsions. *IEEE Trans. Automatic Control*, 56(7):1567 –1581, July 2011.

[50] M. Beer, S. Ferson, and V. Kreinovich. Imprecise probabilities in engineering analyses. *Mechanical Systems and Signal Processing*, 37(1):4 – 29, 2013.

[51] B. Fortin, S. Hachour, and F. Delmotte. Multi-target PHD tracking and classification using imprecise likelihoods. *Intern. Journal of Approximate Reasoning*, 90:17 – 36, 2017.

[52] T. Augustin, F. Coolen, g. de Cooman, and M. Troffaes. *Introduction to imprecise probabilities.* John Wiley & Sons, 2014.

[53] E. Miranda. A survey of the theory of coherent lower previsions. *Intern. Journal of Approximate Reasoning*, 48(2):628 – 658, 2008.

[54] F. Cuzzolin. Credal semantics of Bayesian transformations in terms of probability intervals. *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(2):421–432, 2010.

[55] A. Antonucci, C. P. de Campos, D. Huber, and M. Zaffalon. Approximate credal network updating by linear programming with applications to decision making. *Intern. Journal of Approximate Reasoning*, 58:25–38, 2015.

[56] M. C. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.

[57] P. Walley. Measures of uncertainty in expert systems. *Artificial intelligence*, 83(1):1–58, 1996.

[58] M. Troffaes. Uncertainty and conflict: A behavioural approach to the aggregation of expert opinions. In *Proceedings of the 6th Workshop on Uncertainty Processing*, pages 263–277, 2003.

[59] A. Benavoli and A. Antonucci. An aggregation framework based on coherent lower previsions: Application to zadehs paradox and sensor networks. *International journal of approximate reasoning*, 51(9):1014–1028, 2010.

[60] A. Antonucci and D. Huber. G-LP an algorithm for approximated credal network inferences. `http://ipg.idsia.ch/software.php?id=135`, since 2013.