Partial Classification in the Belief Function Framework

Liyao Ma, Thierry Denoeux

Partial Classification in the Belief Function Framework

Liyao Ma[a], Thierry Denœux[b,c,d]

[a] *University of Jinan, Jinan, China*
[b] *Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France*
[c] *Shanghai University, UTSEUS, Shanghai, China*
[d] *Institut universitaire de France, Paris, France*

Abstract

Partial, or set-valued classification assigns instances to sets of classes, making it possible to reduce the probability of misclassification while still providing useful information. This paper reviews approaches to partial classification based on the Dempster-Shafer theory of belief functions. To define the utility of set-valued predictions, we propose to extend the utility matrix using an Ordered Weighted Average operator, allowing us to model the decision maker's attitude towards imprecision using a single parameter. Various decision criteria are analyzed comprehensively. In particular, two main strategies are distinguished: partial classification based on complete preorders among partial assignments, and partial preorders among complete assignments. Experiments with UCI and simulated Gaussian data sets show the superiority of partial classification in terms of average utility, as compared to single-class assignment and classification with rejection.

Keywords: Dempster-Shafer theory, evidence theory, supervised classification, decision-making, set-valued classification, OWA operator

1. Introduction

In machine learning, classification involves identifying which category a new observation belongs to, based on a training set. Traditionally, when learning a classification model, the input space is divided into as many decision regions as classes, separated by decision boundaries. Given a set of n possible labels, an instance is assigned to one and only one of the n classes. However, such a hard partitioning of the input space often leads to misclassification in case of high uncertainty. For example, ambiguity occurs for observations lying near the boundaries of decision regions, where multiple classes have similar probabilities. Also, when the training set is small, the estimated decision boundaries may significantly differ from the optimal ones, resulting in poor classifier reliability.

Rejection is a classical way to deal with this problem [4][22][20]. Basically, it consists in abstaining from making a decision (i.e., assigning the pattern to a class) when the uncertainty

Email addresses: cse_maly@ujn.edu.cn (Liyao Ma), thierry.denoeux@utc.fr (Thierry Denœux)

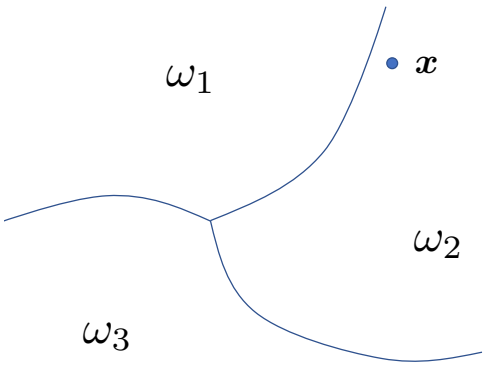


Figure 1: A situation of classification uncertainty with three decision regions. Pattern $\boldsymbol{x}$ cannot be reliably assigned to classes ω_1 or ω_2, but it almost certainly does not belong to ω_3. The set-valued prediction $\{\omega_1, \omega_2\}$ is a reliable option in this case.

too high, making it possible to reduce the probability of misclassification. However, there are cases where, although the risk of misclassification is high, a *subset* of labels can still be considered as very plausible. For instance, in the example illustrated in Figure 1, pattern $\boldsymbol{x}$ cannot be classified with high certainty, but it almost certainly does not belong to class ω_3. In such a case, assigning $\boldsymbol{x}$ to the set of classes $\{\omega_1, \omega_2\}$ seems to be a more reasonable option than outright rejection. Such *partial*[1], or *set-valued classification* makes it possible to better reflect uncertainty and increase the reliability of classifiers. Classification with a reject option can be seen as a special case of partial classification, rejection being equivalent to assigning the pattern to the entire set of possible labels. Compared to rejection, partial classification makes it possible to provide more informative decisions while still minimizing the risk of misclassification.

In this paper, we propose to tackle partial classification in the framework of the *Dempster-Shafer (DS) theory of belief functions* [7][37][15][11], a general framework for reasoning and making decisions under uncertainty[2]. Classifiers based on DS theory, called *evidential classifiers*, quantify prediction uncertainty using belief functions, which provide a more flexible representation of uncertainty than probabilities [8][10][14]. In the early days of DS theory, it was not clear how to make rational decisions based on belief functions, but many approaches have been proposed in the last 20 years, as recently reviewed in [13]. The aim of this paper is to carry out a systematic theoretical and experimental investigation of decision rules for partial classification in the belief function framework.

An important step to implement partial classification is to define the value, or *utility* of set-valued predictions. For instance, consider a three-class problem. If the utility of correct

[1]In this paper, "partial classification" refers to making set-valued predictions in classification tasks. It should not be confused with partial classification (also called "nugget discovery") [1, 34] in the context of descriptive knowledge discovery, which involves mining rules for target classes of interest.

[2]This paper is a revised and extended version of the short paper [30] presented at the ISIPTA 2019 conference, Ghent, Belgium, 3-6 July, 2019.

classification is 1 and the utility of misclassification is 0, what is the utility of predicting the set $\{\omega_1, \omega_2\}$ if the true class is, say, ω_1? It should be strictly greater than 0, as the classifier correctly eliminated class ω_3, but strictly less than 1, as a correct but imprecise prediction is arguably less valuable than a prediction that is both correct and precise. Here, we propose a method to define the utilities of set-valued predictions based on the utility of precise ones and a single control parameter, using Ordered Weighted Average (OWA) operators [46]. Having defined an extended utility matrix, a generalized Maximum Expected Utility (MEU) principle allows us to make set-valued predictions based on different notions of expectation with respect to a belief function. In the following, we briefly review previous approaches to partial classification.

Related work

Learning set-valued classifiers has been implemented with various approaches. Vovk et al. [44] propose *conformal prediction*, an approach for learning set-valued classifiers with finite sample confidence guarantees. In the same vein, Sadinle et al. [35] design classifiers that guarantee user-defined levels of coverage while minimizing ambiguity. This research direction is representative of the statistical approach, in which set-valued predictions are viewed as generalizations of confidence or credible intervals.

Another research direction, more rooted in decision theory, involves utility (or, equivalently, loss) functions. In the probabilistic framework, Ha [21] introduces a simple model in which the loss of making a set-valued prediction is the sum of two terms, one reflecting the loss of missing the true class, and the other one penalizing imprecision. He then proposes an efficient algorithm to minimize the expected loss with respect to conditional class probabilities. Mortier et al. [31] also assume uncertainty to be quantified by conditional class probabilities, and the quality of a predicted set to be measured by a utility function. They, then, address the problem of finding the Bayes-optimal prediction, i.e., the subset of class labels with highest expected utility. Following a different approach, del Coz et al. [6] make a parallel between partial (or "nondeterministic") classification and information retrieval, and propose a loss function inspired by the F_β measure for aggregating precision and recall. In this utility-based approach, the definition of the utility (or loss) of a set-valued prediction is essential. Yang et al. [49] list some properties losses of set-valued predictions should meet, and they measure the loss of predicting a set of classes as a generalized mean of the losses of predicting each individual class in that set.

In the imprecise probability framework, Zaffalon [50] introduces the naive credal classifier, in which prior ignorance about the distribution of classes is modeled by means of a set of prior densities (also called the prior *credal set*). This credal set is turned into a set of posterior probabilities by element-wise application of Bayes' rule. The classifier returns all the classes that are non-dominated by any other class according to the posterior credal set. As we will see in Section 3.4, a similar approach can be implemented with evidential classifiers. Zaffalon et al. [51] propose a metric to evaluate the predictions of credal classifiers, considering the $\{0, 1\}$ reward case and taking the decision maker's degree of risk aversion into account.

In the DS framework, most authors only consider precise assignment to the class with maximum plausibility [2][16] or maximum pignistic probability [39][33][42]. Denœux [9] pro-

poses several decision rules based on the maximization of upper, lower or pignistic expected utility; however, he only considers precise class assignment and rejection. Several authors propose heuristic decision strategies for partial classification. As an evidential classifier typically outputs a mass function, i.e., a mapping that assigns each set of classes a number between 0 and 1, a simple approach is to select the set of classes with maximal mass [29]. This approach, however, can be criticized because the mass assigned to a subset does not measure its credibility or plausibility. For instance, in a four-class problem, if we have a mass 0.4 on the set composed of classes ω_1 and ω_2, a mass 0.3 on class ω_3, and a mass 0.3 on class ω_4, it would be paradoxical to select the set $\{\omega_1, \omega_2\}$, as the set $\{\omega_3, \omega_4\}$ is actually more supported by the evidence. Other authors develop more sophisticated strategies. For instance, Liu et al. [28] propose to select the classes ω such that the pignistic probability $p(\omega)$ is larger than some constant ϵ times the maximum pignistic probability. Parameter ϵ is tuned to maximize a "benefit value" that depends on the cardinality of the sets. Liu et al. use a similar approach in [27] but implement a strategy that selects either a single class, or a pair of classes.

From the above review of related work, it appears that no systematic study of principled decision-theoretic approaches to partial classification in the Dempster-Shafer framework has been undertaken so far, a gap that we aim to fill in this work. The rest of this paper is organized as follows. Section 2 makes the paper self-contained by providing a brief reminder of basic definitions and notations used later on. Our approach is introduced in Section 3: after proposing a method for defining the utility of set-valued predictions, we review decision criteria for partial classification, and we address the evaluation of classification performance. Extensive experimental comparisons of different decision criteria are then presented in Section 4 using UCI and artificial Gaussian data sets. Finally, the main conclusions are summarized in Section 5.

2. Background

In this section, we review background notions and define the notations. The theory of belief functions is first reviewed in Section 2.1 and the OWA operators are recalled in Section 2.2. The decision framework is defined in Section 2.3.

2.1. Theory of belief functions

As a generalization of both set and probabilistic uncertainty, the theory of belief functions [7, 37] provides a general framework for modelling and reasoning with uncertainty. Here only a few definitions needed in the rest of the paper are recalled. More complete descriptions can be found in Shafer's book [37] and in the recent survey [15].

Let $\Omega = \{\omega_1, \cdots, \omega_n\}$ be a finite set, called the *frame of discernment*, assumed to contain all the possible exclusive values that a variable (in the classification problem, the label of an instance) can take. When the true value of the variable is ill-known, partial information about it can be modeled by a *mass function* $m : 2^\Omega \to [0, 1]$ such that $m(\emptyset) = 0$ and

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

A subset A of Ω with positive mass is called a *focal set* of m. The quantity $m(A)$ can then be interpreted as the amount of evidence indicating that the true value is specifically in A (and in no strict subset). This formalism extends both probabilistic and set-valued uncertainty models. In particular, the *vacuous* mass function verifies $m(\Omega) = 1$ and represents total ignorance. A *Bayesian* mass function is such that all its focal sets are singletons; it is equivalent to a probability distribution. A mass function such that $m(A) = 1$ for some subset $A \subseteq \Omega$ is said to be *logical*; it is equivalent to A.

The *belief* and *plausibility* functions corresponding to mass function m are defined, respectively, as

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2}$$

and

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\overline{A}), \tag{3}$$

for all $A \subseteq \Omega$. The belief function sums up the masses assigned to subsets of A and measures how much event A is supported by the evidence, while the plausibility measures how much event A is consistent with the evidence. These two functions define the lower and upper bounds of the set $\mathcal{P}$ of probability measures P *compatible* with mass m, i.e, such that $Bel(A) \leq P(A) \leq Pl(A)$ for all $A \subseteq \Omega$.

Two mass functions m_1 and m_2 representing independent items of evidence can be combined using *Dempster's rule* [37] as follows,

$$(m_1 \oplus m_2)(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B) m_2(C)}, \tag{4}$$

for all $A \subseteq \Omega, A \neq \emptyset$ and $(m_1 \oplus m_2)(\emptyset) = 0$. Dempster's rule is commutative and associative.

In the Transferable Belief Model (TBM) [40], Smets proposed to make decisions based on the *pignistic probability distribution* [39], which is obtained by distributing masses equally among the sets of A,

$$BetP(\omega) = \sum_{\{A \subseteq \Omega : \omega \in A\}} \frac{m(A)}{|A|}, \tag{5}$$

where $|A|$ denotes the cardinality of $A \subseteq \Omega$. Other decision criteria are reviewed in [13].

2.2. Ordered Weighted Average operators

The OWA operators proposed by Yager [46] are a parametrized class of mean type aggregation operators, including common operators such as the maximum, the arithmetic average and the minimum. An OWA operator of dimension n is formally defined as a mapping $F_{\boldsymbol{w}}$ from $\mathbb{R}^n$ to $\mathbb{R}$ with associated collection of positive weights $\boldsymbol{w} = (w_1, \cdots, w_n)$ summing up to one, such that

$$F_{\boldsymbol{w}}(a_1, \cdots, a_n) = \sum_{i=1}^{n} w_i b_i, \tag{6}$$

where b_i is the i-th largest element in $a_1, \cdots, a_n$. Different aggregating operators can be implemented by using different weights. Yager [46] defined the measure of *orness* as

$$\mathsf{orness}(\boldsymbol{w}) = \frac{1}{n-1}\sum_{i=1}^{n}(n-i)w_i, \tag{7}$$

which describes the behavior of the operator. The maximum, the arithmetic average and the minimum correspond, respectively, to orness measures of 1, 0.5 and 0.

O'Hagan [32] proposed to determine the weights by fixing the orness measure to some value $\gamma \in [0, 1]$, and searching for the weight vector $\boldsymbol{w}^*$ that maximizes the entropy

$$H(\boldsymbol{w}) = -\sum_{i=1}^{n} w_i \log w_i \tag{8}$$

under the constraint $\mathsf{orness}(\boldsymbol{w}) = \gamma$. Filev and Yager [18] showed that the optimal weights $w_1^*, \ldots, w_n^*$ form a geometric sequence, which is strictly decreasing if $\gamma > 0.5$ and strictly increasing if $\gamma < 0.5$. Liu and Chen [26] showed that, for all $(a_1, \ldots, a_n) \in \mathbb{R}^n$,

$$F_{\boldsymbol{w}^*}(a_1, \cdots, a_n) \geq \frac{1}{n}\sum_{i=1}^{n} a_i \tag{9a}$$

if $\gamma \geq 0.5$ and

$$F_{\boldsymbol{w}^*}(a_1, \cdots, a_n) \leq \frac{1}{n}\sum_{i=1}^{n} a_i \tag{9b}$$

if $\gamma \leq 0.5$.

2.3. Decision-making framework

The purpose of classification is to build a model (called a *classifier*) that maps feature vectors (or attributes) to class labels representing the object category. Once a classifier has been trained, it is used to make predictions for new instances whose classes are unknown. Throughout the paper, we will consider the common model of finite decision theory, i.e., we assume that the set $\Omega = \{\omega_1, \cdots, \omega_n\}$ of classes (or "states of nature" using the terminology of decision theory) is finite. This set will constitute the frame of discernment on which belief functions will be defined.

To make decisions on label prediction (instance assignment), the decision maker (DM) needs to choose an *act* f from a finite set $\mathcal{F}$. Generally, precise predictions are required, so that we consider only acts assigning an instance to one and only one of the n classes $\omega_i \in \Omega$; such acts are called *precise assignments*. The set of available acts is then a finite set containing n elements, denoted as $\mathcal{F} = \{f_1, \cdots, f_n\}$, where f_i represents the act of assigning the instance to class ω_i. Formally, an act is defined as a mapping from Ω to a set $\mathscr{C}$ of consequences (or outcomes). In classification problems, consequences are of the form c_{ij}, defined as the consequence of assigning an instance to class ω_i when it actually belongs to class ω_j. Therefore, act f_i maps each state ω_j to consequence c_{ij}.

To measure the desirability of consequences, we usually define a utility function U from $\mathscr{C}$ to $[0,1]$. The utilities $u_{ij} = U(c_{ij})$ can be arranged in a *utility matrix* $\mathbf{U}$ of size $n \times n$. The general term u_{ij} of $\mathbf{U}$ represents the utility of selecting act f_i (predicting the label as ω_i) when the true class is ω_j. In practice, matrix $\mathbf{U}$ is often assumed to be the identity matrix, in which case the utility is equal to 1 for a correct decision, and 0 for a misclassification. However, the DM can also assign different utilities to different types of misclassification, making $\mathbf{U}$ different from the identity matrix, and possibly asymmetric[3].

Given the set $\mathcal{F}$ of acts, utility matrix $\mathbf{U}$, and some measure of uncertainty over Ω, the DM's preference over acts is denoted by $\succcurlyeq$, where $f \succcurlyeq g$ means that act f is at least as desirable as g. The strict preference and indifference relations are denoted in the usual manner, respectively, as $f \succ g$ (f is strictly more desirable than g) and $f \sim g$ (f and g are equally desirable). Relation $\succcurlyeq$ is usually assumed to be reflexive (for any f, $f \succcurlyeq f$) and transitive (for any f, g, h, if $f \succcurlyeq g$ and $g \succcurlyeq h$, then $f \succcurlyeq h$). A reflexive and transitive preference relation is a *preorder*. If, furthermore, $\succcurlyeq$ is antisymmetric (for any f, g, if $f \succcurlyeq g$ and $g \succcurlyeq f$, then $f = g$), then it is an order. If for any two acts f and g, $f \succcurlyeq g$ or $g \succcurlyeq f$, the preference relation is *complete*, otherwise, it is *partial*. Most decision rules induce a complete or partial preorder on $\mathcal{F}$. An act f is a *greatest element* of relation $\succcurlyeq$ if it is at least as desirable as any other act, *i.e.*, for any $f' \in \mathcal{F}$, $f \succcurlyeq f'$. A complete preorder always has at least one greatest element, and it has only one if it is a complete order. An act f is a *maximal (or non-dominated) element* of the strict preference relation if no other act is strictly preferred to f, i.e., if for any $f' \in \mathcal{F}$, $\neg(f' \succ f)$. A greatest element is a maximal element, but the converse is not true in general [13].

In the case of *partial classification*, prediction is not limited to precise labels, but can take the form of any non-empty subset of Ω. The act of assigning an instance to a subset K of Ω (with cardinality $|K| > 1$) is called a *partial assignment* and is denoted as f_K. To achieve a preorder among available acts $\widetilde{\mathcal{F}} = \{f_K, K \subseteq \Omega, K \neq \emptyset\}$, the utility for each set-valued prediction must be defined. This problem is addressed in the next section.

3. Decision-making for evidential classification

Assuming the DM's information concerning the possible states of nature to be represented by a mass function m on Ω, we now carry out a thorough analysis and comparison of different decision-making criteria to obtain label predictions, especially set-valued ones. A method for computing the utility of set-valued predictions is first introduced in Section 3.1. Precise classification with and without rejection is then recalled in Section 3.2, after which two main approaches to partial classification are studied: via complete preorder among partial assignments (Section 3.3) and via partial preorder among complete assignments (Section 3.4). In Section 3.5, the time complexity issue is considered. Finally, the evaluation of set-based predictions is discussed in Section 3.6.

[3]Especially in cost-sensitive problems such as ordinal classification [19] and imbalanced classification [23], where different prediction errors are assumed to be treated differently.

3.1. Generating utilities of partial assignments via an OWA operator

Given the set of classes $\Omega = \{\omega_1, \cdots, \omega_n\}$, we have a utility matrix $\mathbf{U} = (u_{ij})_{n\times n}$ specifying the utilities of precise assignments. Without loss of generality, we assume the utilities to be defined on the scale $[0, 1]$, with the diagonal terms of $\mathbf{U}$ equal to 1 (the assignment to the correct class has maximum utility). When partial assignments are considered, $\mathbf{U}$ must be extended to a $(2^n - 1) \times n$ matrix $\widetilde{\mathbf{U}}$, whose general term $\widetilde{u}_{K,j}$ represents the utility of assigning an instance to the set K of possible classes when the true class is ω_j. In some applications, it might be possible to elicit the extended matrix by asking the DM to assess utilities of set-valued predictions. In the following, we discuss a general way to construct the extended utility matrix $\widetilde{\mathbf{U}}$ directly from the original one $\mathbf{U}$ of size $n \times n$ using a single parameter.

Before describing our proposal, we start with a discussion on the extended utilities. Intuitively, given a state of nature ω_j, the utility $\widetilde{u}_{K,j}$ of assigning an instance to set K should be a function of the utilities of each precise assignments within the set (*i.e.*, utilities u_{ij} such that $\omega_i \in K$). A DM totally indifferent to imprecision would set $\widetilde{u}_{K,j} = \max_{\omega_i \in K} u_{ij}$. In this case, as long as the true label is included in K, the partial assignment f_K achieves utility 1 no matter how imprecise K is, so that imprecision is not penalized. A more imprecision-averse attitude would be to define the utility of a partial assignment as the average of utilities of precise assignments within that set, i.e., $\widetilde{u}_{K,j} = \overline{u}_{K,j}$ with

$$\overline{u}_{K,j} = \frac{1}{|K|} \sum_{\omega_i \in K} u_{ij}.$$

We note that $\overline{u}_{K,j}$ is equal to the expected utility of picking one label uniformly at random from set K. It would be irrational to set $\widetilde{u}_{K,j}$ to a lower value, because given a set K of labels, we can always pick one label at random and adopt it as our precise assignment, in which case the expected utility would be equal to $\overline{u}_{K,j}$. In general, we can thus reasonably assume the following inequalities to hold:

$$\overline{u}_{K,j} \leq \widetilde{u}_{K,j} \leq \max_{\omega_i \in K} u_{ij} \tag{10}$$

for all non empty subset K of Ω and all $\omega_j \in \Omega$.

As a general model, we can further assume $\widetilde{u}_{K,j}$ to result from the aggregation of utilities u_{ij} for states ω_i in K using some OWA operator (see Section 2.2) with weight vector $\boldsymbol{w}$ of length $|K|$:

$$\widetilde{u}_{K,j} = F_{\boldsymbol{w}}\left(\{u_{ij} : \omega_i \in K\}\right) = \sum_{k=1}^{|K|} w_k u^K_{(k)j}, \tag{11}$$

where $u^K_{(k)j}$ denotes the k-th largest element in the set $\{u_{ij},\ \omega_i \in K\}$. In (11), each weight w_k can be interpreted as measuring the DM's preference to choose $u^K_{(k)j}$ if forced to select a single value in $\{u_{ij} : \omega_i \in K\}$. Similar to Yager's definition of the orness degree (7), we

Table 1: Utility matrices $\widetilde{\mathbf{U}}$ extended by OWA operators with $\gamma = 0.8$ and $\gamma = 0.6$ (Example 1).

act	$\gamma = 0.8$			$\gamma = 0.6$		
	ω_1	ω_2	ω_3	ω_1	ω_2	ω_3
$f_{\{\omega_1\}}$	1.0000	0.2000	0.1000	1.0000	0.2000	0.1000
$f_{\{\omega_2\}}$	0.2000	1.0000	0.2000	0.2000	1.0000	0.2000
$f_{\{\omega_3\}}$	0.1000	0.2000	1.0000	0.1000	0.2000	1.0000
$f_{\{\omega_1,\omega_2\}}$	0.8400	0.8400	0.1800	0.6800	0.6800	0.1600
$f_{\{\omega_1,\omega_3\}}$	0.8200	0.2800	0.8200	0.6400	0.2000	0.6400
$f_{\{\omega_2,\omega_3\}}$	0.1800	0.8400	0.8400	0.1600	0.6800	0.6800
$f_{\{\omega_1,\omega_2,\omega_3\}}$	0.7373	0.7455	0.7373	0.5269	0.5507	0.5269

define the DM's *imprecision tolerance degree* as

$$\mathsf{tol}(\boldsymbol{w}) = \sum_{k=1}^{|K|} \frac{|K| - k}{|K| - 1} w_k = \gamma. \tag{12}$$

Given γ, the weights corresponds to the OWA operator can be obtained by maximizing the entropy (8), subject to $\mathsf{tol}(\boldsymbol{w}) = \gamma$ and $\sum_{k=1}^{|K|} w_k = 1$.

The OWA-based approach makes it possible to parameterize the DM's tolerance to imprecision by a single parameter γ. The higher the value of γ, the more imprecision is tolerated. Setting $\gamma = 1$ gives us the maximum operator, and $\gamma = 0.5$ gives us the average. From (9), setting γ in the range $[0.5, 1]$ is a necessary and sufficient condition for the equalities (10) to be satisfied. Example 1 below illustrates the process of aggregating utilities via OWA operators.

Example 1. *Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and consider the utility matrix*

$$\mathbf{U} = \begin{bmatrix} 1 & 0.2 & 0.1 \\ 0.2 & 1 & 0.2 \\ 0.1 & 0.2 & 1 \end{bmatrix}.$$

Assuming that the true label is ω_1, Figure 2 shows the aggregated utilities for sets $\{\omega_1, \omega_2\}$, $\{\omega_1, \omega_3\}$ and $\{\omega_1, \omega_2, \omega_3\}$ for different values of γ. The aggregated utility is only related to the utilities of elements within the set. As γ ranges from 0 to 1, the extended utility for each set varies from the minimal utility in this set to the maximal one. When $\gamma = 0.5$, the average utility is obtained. As mentioned above, we only need to consider values of $\gamma > 0.5$, since when $\gamma \leq 0.5$ precise predictions are always more preferable. Table 1 shows the extended utility matrices obtained by OWA operators with $\gamma = 0.8$ and $\gamma = 0.6$.

3.2. Precise classification with and without rejection

Given utility matrix $\mathbf{U}$, precise predictions are based on a complete preorder among precise assignments in $\mathcal{F} = \{f_1, \ldots, f_n\}$. In [9], Denœux considered the following three

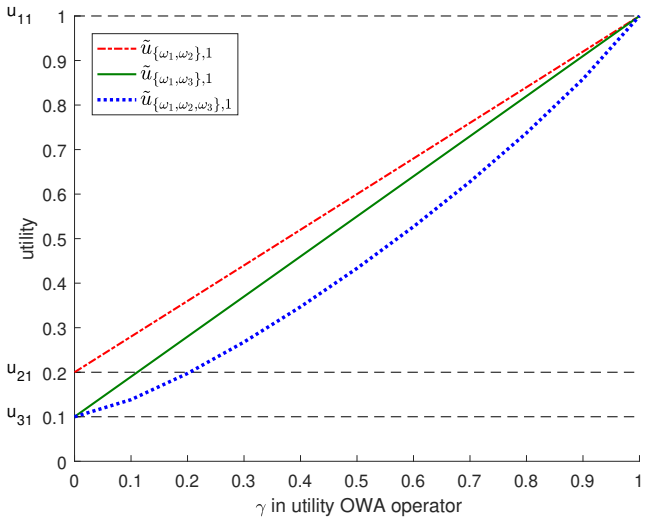


Figure 2: Aggregated utilities vs. imprecision tolerance degree γ.

decision criteria:

Generalized maximin criterion. The complete preorder among precise assignments is defined by the lower expected utility

$$\underline{\mathbb{E}}_m(f_i) = \sum_{B\subseteq\Omega} m(B) \min_{\omega_j\in B} u_{ij}, \tag{13}$$

for all $f_i \in \mathcal{F}$. For two arbitrary precise assignments, we obtain the preference relation $f_i \succcurlyeq_* f_j \iff \underline{\mathbb{E}}_m(f_i) \geq \underline{\mathbb{E}}_m(f_j)$. Relation $\succcurlyeq_*$ corresponds to a pessimistic attitude of the DM, as he considers the least desirable consequence within each focal set B.

Generalized maximax criterion. Taking an optimistic point of view, a complete preorder can be defined from the upper expected utility

$$\overline{\mathbb{E}}_m(f_i) = \sum_{B\subseteq\Omega} m(B) \max_{\omega_j\in B} u_{ij}, \tag{14}$$

with $f_i \succcurlyeq^* f_j \iff \overline{\mathbb{E}}_m(f_i) \geq \overline{\mathbb{E}}_m(f_j)$. This criterion reflects an optimistic, or ambiguity-seeking attitude of the DM.

Pignistic criterion. This criterion averages the utilities within each focal set. The act with higher average utility will be more preferred, which means that $f_i \succcurlyeq_p f_j \iff \mathbb{E}_p(f_i) \geq \mathbb{E}_p(f_j)$, where

$$\mathbb{E}_p(f_i) = \sum_{i=1}^{n} BetP(\omega_j)u_{ij} = \sum_{i=1}^{n} \left(\sum_{B\ni\omega_j} \frac{m(B)}{|B|} \right) u_{ij} = \sum_{B\subseteq\Omega} m(B) \left(\frac{1}{|B|} \sum_{\omega_j\in B} u_{ij} \right). \tag{15}$$

In [9], Denœux also proposed decision rules with rejection. Instances likely to be misclassified are rejected. Rejection can be identified with an additional act f_Ω that consists in assigning the instance to the whole set of classes. In [9], it was proposed to set $u_{\Omega,j}$ to some fixed value λ_0 for all j.

In the next two sections, we extend the notion of rejection by allowing partial assignment not only to Ω, but also to any subset of Ω.

3.3. Partial classification via complete preorders among partial assignments

As explained in Section 2.3, when considering partial assignments, the set of available acts is $\widetilde{\mathcal{F}} = \{f_K, K \subseteq \Omega, K \neq \emptyset\}$. A mass function m on Ω and an extended utility matrix $\widetilde{\mathbf{U}}_{(2^n-1)\times n}$ constructed, e.g., using the approach described in Section 3.1, are used for decision-making. In addition to the previous three decision criteria, the following additional criteria reviewed in [13] will be investigated:

Generalized Hurwicz criterion. This criterion considers a convex combination of the minimum and maximum utility, with a *pessimism index* $\alpha \in [0, 1]$ adjusting the combination [41]. The generalized maximin and maximax criteria are special cases corresponding, respectively, to $\alpha = 1$ and $\alpha = 0$. For two acts f_K and f_G corresponding to nonempty subsets of classes K and G, we have $f_K \succcurlyeq_\alpha f_G \iff \mathbb{E}_{m,\alpha}(f_K) \geq \mathbb{E}_{m,\alpha}(f_G)$ with

$$\begin{aligned}\mathbb{E}_{m,\alpha}(f_K) &= \sum_{B\subseteq\Omega} m(B)\left(\alpha \min_{\omega_j\in B} \widetilde{u}_{K,j} + (1-\alpha)\max_{\omega_j\in B} \widetilde{u}_{K,j}\right)\\ &= \alpha\underline{\mathbb{E}}_m(f_K) + (1-\alpha)\overline{\mathbb{E}}_m(f_K). \end{aligned} \tag{16}$$

Generalized OWA criterion. Another generalization of the maximin and maximax criteria consists in aggregating the utilities within each focal set $K \subseteq \Omega$ using OWA operators [47]. We have

$$\mathbb{E}^{owa}_{m,\beta}(f_K) = \sum_{B\subseteq\Omega} m(B) F_\beta(\{\widetilde{u}_{K,j} : \omega_j \in B\}), \tag{17}$$

where F_β is the maximum entropy OWA operator with degree of orness (or optimism) β. We have $f_K \succcurlyeq_\beta f_G \iff \mathbb{E}^{owa}_{m,\beta}(f_K) \geq \mathbb{E}^{owa}_{m,\beta}(f_G)$. The pignistic criterion is recovered when $\beta = 0.5$.

Generalized minimax regret criterion This criterion extends Savage's minimax regret criterion [36] to decision-making with belief functions [48]. Defining the regret that act f_K is selected when the true state ω_j occurs as $r_{K,j} = \max_G \widetilde{u}_{G,j} - \widetilde{u}_{K,j}$, the expected maximal regret for act f_K is

$$\overline{R}(f_K) = \sum_{B\subseteq\Omega} m(B) \max_{w_j\in B} r_{K,j}. \tag{18}$$

For two partial assignments f_G and f_K, we have $f_K \succcurlyeq_r f_G \iff \overline{R}(f_K) \leq \overline{R}(f_G)$.

Table 2: Extended utility matrix and expected utilities for Example 2.

acts	ω_1	ω_2	$\underline{\mathbb{E}}_m(f_K)$	$\overline{\mathbb{E}}_m(f_K)$	$\mathbb{E}_p(f_K)$	$\mathbb{E}_{m,\alpha}(f_K)$	$\mathbb{E}^{owa}_{m,\beta}(f_K)$	$\overline{R}_K$
$f_{\{\omega_1\}}$	1	0.2	0.7600	0.9200	0.8400	0.8080	0.8880	0.2400
$f_{\{\omega_2\}}$	0.3	1	0.3700	0.5100	0.4400	0.4120	0.4820	0.6300
$f_{\{\omega_1,\omega_2\}}$	0.79	0.76	0.7810	0.7870	0.7840	0.7828	0.7858	0.2190

Given a decision criterion, ranking the acts according to their expected utility yields a complete preorder. The best acts are the greatest elements of this preorder. Usually, there is a unique greatest element f_{K^*}, which predicts that the class of the instance belongs to set of labels K^*. It is remarkable that this approach can produce set-valued predictions even with precise probabilities of states of nature. The MEU principle works as a special case to provide complete preorder among partial assignments when uncertainty about the decision is captured by probabilities $p_1, \cdots, p_n$ on Ω rather than a mass function m. The act with greatest expected utility is then the most desirable: $f_K \succcurlyeq_u f_G \iff EU(f_K) \geq EU(f_G)$, where $EU(f_K) = \sum_{j=1}^{n} \widetilde{u}_{K,j} p_j$.

Example 2. *To develop an intuitive understanding of the proposed approach, consider the simple case of binary classification with* $\Omega = \{\omega_1, \omega_2\}$. *Given the asymmetric utility matrix*

$$\mathbf{U} = \begin{bmatrix} 1 & 0.2 \\ 0.3 & 1 \end{bmatrix},$$

the utility matrix $\widetilde{\mathbf{U}}$ *extended by an OWA operator with* $\gamma = 0.7$ *and expected utilities calculated with mass function* $m(\omega_1) = 0.7,\ m(\omega_2) = 0.1,\ m(\{\omega_1, \omega_2\}) = 0.2$ *are summarized in Table 2. According to the results, different decision criteria yield the following strict preference relations:*

- *Generalized maximin criterion:* $f_{\{\omega_1,\omega_2\}} \succ_* f_{\{\omega_1\}} \succ_* f_{\{\omega_2\}}$
- *Generalized maximax criterion:* $f_{\{\omega_1\}} \succ^* f_{\{\omega_1,\omega_2\}} \succ^* f_{\{\omega_2\}}$
- *Pignistic criterion:* $f_{\{\omega_1\}} \succ_p f_{\{\omega_1,\omega_2\}} \succ_p f_{\{\omega_2\}}$
- *Generalized Hurwicz criterion (*$\alpha = 0.7$*):* $f_{\{\omega_1\}} \succ_\alpha f_{\{\omega_1,\omega_2\}} \succ_\alpha f_{\{\omega_2\}}$
- *Generalized OWA criterion (*$\beta = 0.8$*):* $f_{\{\omega_1\}} \succ_\beta f_{\{\omega_1,\omega_2\}} \succ_\beta f_{\{\omega_2\}}$
- *Generalized minimax regret criterion:* $f_{\{\omega_1,\omega_2\}} \succ_r f_{\{\omega_1\}} \succ_r f_{\{\omega_2\}}$

It can be seen that with the same mass function, various criteria yield different predictions.

It is also worthwhile to notice the general relation between the generalized maximin and minimax regret criteria shown in Proposition 1.

Proposition 1. *Given a set* Ω *of classes, a mass function* m *on* Ω*, a utility matrix* $\mathbf{U}$ *and its extension* $\widetilde{\mathbf{U}}$*, the maximin and minimax regret criteria always yield the same complete preorder of partial assignments, as long as correct classifications have the same utility value in* $\mathbf{U}$.

Proof. Assume that all the correct precise predictions have the maximum utility M. We have $\max_{\emptyset \neq G \subseteq \Omega} \widetilde{u}_{G,j} = M$. Then,

$$\begin{aligned}
\underline{\mathbb{E}}_m(f_K) + \overline{R}(f_K) &= \sum_{\emptyset \neq B \subseteq \Omega} m(B) \min_{\omega_j \in B} \widetilde{u}_{K,j} + \sum_{\emptyset \neq B \subseteq \Omega} m(B) \max_{\omega_j \in B} r_K(\omega_j) \\
&= \sum_{\emptyset \neq B \subseteq \Omega} m(B) \min_{\omega_j \in B} \widetilde{u}_{K,j} + \sum_{\emptyset \neq B \subseteq \Omega} m(B) \max_{\omega_j \in B} \left(\max_{\emptyset \neq G \subseteq \Omega} \widetilde{u}_{G,j} - \widetilde{u}_{K,j} \right) \\
&= \sum_{\emptyset \neq B \subseteq \Omega} m(B) \left[\min_{\omega_j \in B} \widetilde{u}_{K,j} + \max_{\omega_j \in B} (M - \widetilde{u}_{K,j}) \right] \\
&= \sum_{\emptyset \neq B \subseteq \Omega} m(B) \left[\min_{\omega_j \in B} \widetilde{u}_{K,j} + M - \min_{\omega_j \in B} \widetilde{u}_{K,j} \right] \\
&= M \sum_{\emptyset \neq B \subseteq \Omega} m(B) = M.
\end{aligned}$$

Since $f_K \succcurlyeq_* f_G \iff \underline{\mathbb{E}}_m(f_K) \geq \underline{\mathbb{E}}_m(f_G)$ and $f_K \succcurlyeq_r f_G \iff \overline{R}(f_K) \leq \overline{R}(f_G)$, the complete preference relations induced by the lower expectation and by the expected maximal regret are identical. Therefore, these two decision criteria always reach the same decision. □

3.4. Partial classification via partial preorders among precise assignments

Set-valued predictions can also be induced by considering partial preorders among precise assignments [13]. In this case, the set of available acts $\mathcal{F} = \{f_1, \cdots, f_n\}$, a mass function m on Ω and the utility matrix $\mathbf{U}_{n \times n}$ are used for decision-making. Based on a partial preorder among acts, the choice operator returns an optimal subset $\mathcal{F}^* \subseteq \mathcal{F}$ consisting all the maximal (non-dominated) elements, leading to a set-valued prediction corresponding to $\mathcal{F}^*$.

The decision-making criteria involved in this approach are more rooted in the imprecise probability view of belief functions [43, 13]. With insufficient information about states of nature, each precise assignment f_i corresponds to an expected utility interval $[\underline{\mathbb{E}}_m(f_i), \overline{\mathbb{E}}_m(f_i)]$, where $\underline{\mathbb{E}}_m(f_i)$ and $\overline{\mathbb{E}}_m(f_i)$ are calculated by Equations (13) and (14). This interval can also be viewed as the range of expectations $\mathbb{E}_P(f)$ with respect to all probability measures P compatible with mass function m. By comparing the acts in $\mathcal{F}$ based on the lower and upper expected utilities, partial preorders of precise assignments are obtained using the following decision criteria. (The reader is referred to [13] for more a detailed presentation and discussion of these criteria).

Strong dominance criterion. This criterion [43] is based on the strong dominance (also called *interval dominance*) relation stating that, for two precise assignments corresponding to ω_i and ω_j, we have $f_i \succcurlyeq_{SD} f_j$ iff $\underline{\mathbb{E}}_m(f_i) \geq \overline{\mathbb{E}}_m(f_j)$. The non-dominated elements then form the choice set $\mathcal{F}^*_{SD} = \{f \in \mathcal{F} : \nexists f' \in \mathcal{F}, \text{s.t.} f \succcurlyeq_{SD} f'\}$. With the strong condition of interval dominance, many pairs of acts may be not comparable, even making $\mathcal{F}^*_{SD} = \mathcal{F}$.

Weak dominance criterion. According to this less conservative criterion [13], $f_i \succcurlyeq_{WD} f_j$ iff $\left(\underline{\mathbb{E}}_m(f_i) \geq \underline{\mathbb{E}}_m(f_j)\right)$ and $\left(\overline{\mathbb{E}}_m(f_i) \geq \overline{\mathbb{E}}_m(f_j)\right)$. It is clear that $f_i \succcurlyeq_{SD} f_j$ implies $f_i \succcurlyeq_{WD} f_j$. The set of non-dominated elements of $\succcurlyeq_{WD}$ is included in that of $\succcurlyeq_{SD}$, *i.e.*, $\mathcal{F}^*_{WD} \subseteq \mathcal{F}^*_{SD}$.

Maximality criterion. Developed in the imprecise probability framework [45], this criterion can also be used in the belief function framework. The preference between two acts is defined as $f_i \succcurlyeq_{max} f_j \iff \underline{\mathbb{E}}_m(f_i - f_j) \geq 0$. We still have $\mathcal{F}^*_{max} \subseteq \mathcal{F}^*_{SD}$, which comes at the price of higher computational costs.

Interval-valued utility criterion. Denoeux and Shenoy [17] define interval-valued utilities and propose a two-coefficient practical model for utility elicitation. For each focal element of m, utility bounds are defined by convex combinations of utilities of its worst and best outcomes:

$$\underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i) = \sum_{B\subseteq\Omega} m(B)\left[\alpha_u \min_{\omega_j\in B} u_{ij} + (1-\alpha_u)\max_{\omega_j\in B} u_{ij}\right], \tag{19a}$$

$$\overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i) = \sum_{B\subseteq\Omega} m(B)\left[\beta_u \min_{\omega_j\in B} u_{ij} + (1-\beta_u)\max_{\omega_j\in B} u_{ij}\right], \tag{19b}$$

where α_u and β_u are two local pessimism indices reflecting the DM's attitude towards ambiguity and indeterminacy, with $0 \leq \beta_u \leq \alpha_u \leq 1$. The preference relation is then defined as

$$f_i \succcurlyeq_{IVU} f_j \iff \left(\underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i) \geq \underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_j)\right) \text{ and } \left(\overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i) \geq \overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_j)\right).$$

These four criteria induce partial preorders of precise assignments, resulting in a choice set $\mathcal{F}^*$ consisting possibly several greatest elements. In contrast, the e-admissibility criterion defines a choice set immediately without defining a partial preorder.

E-admissibility criterion. E-admissibility labels an act optimal when it dominates any other available acts in expectation with respect to every probability measure, a stronger condition than maximality (with respect to at least one probability measure). An act can be decided to be e-admissible or not by solving the following linear programming problem [24]:

$$\min_{\boldsymbol{a},\boldsymbol{p},\boldsymbol{\lambda}} \sum_{\ell\neq i} \lambda_\ell \tag{20a}$$

subject to

$$\sum_{\{k\,:\,\omega_k\in F_j\}} a_{kj} = m(F_j),\ j = 1,\cdots,q \tag{20b}$$

$$a_{kj} \geq 0\ \forall(k,j) : \exists(\omega_k, F_j),\ \omega_k \in F_j \tag{20c}$$

$$p_k = \sum_{j=1}^{q} a_{kj},\ k = 1, \cdots, n \tag{20d}$$

$$\sum_{k=1}^{n} p_k(u_{ik} - u_{\ell k}) + \lambda_\ell \geq 0,\ \ell \neq i \tag{20e}$$

$$\lambda_\ell \geq 0,\ \ell \neq i, \tag{20f}$$

where $F_1, \cdots, F_q$ are the focal elements of m, vector $\boldsymbol{a}$ contains all the allocations[4] $a_{kj} = a(\omega_k, F_j)$ such that $\omega_k \in F_j$, $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_n)$ are $n-1$ slack variable vectors, and the vector $\boldsymbol{p} = (p_1, \cdots, p_n)$ corresponds to the compatible probabilities. We get the solution $\boldsymbol{\lambda} = 0$ iff act f_i is e-admissible. Obviously, this criterion is much more costly to compute than the others. Thanks to the set relation of different choice sets $\mathcal{F}^*_{ead} \subseteq \mathcal{F}^*_{max} \subseteq \mathcal{F}^*_{SD}$, to reduce computational cost, we can solve the linear program (20) only for those acts within the choice set of the Maximality criterion.

All the criteria reviewed in this section provide a choice set $\mathcal{F}^* = \{f \in \mathcal{F} : \forall f' \in \mathcal{F}, \neg(f' \succ f)\} \subseteq \mathcal{F}$, which contains the non-dominated precise assignments. The acts within set $\mathcal{F}^*$ are not comparable, making the prediction set-valued, as illustrated in Example 3. In general, partial preorders result from lack of information. With sufficient knowledge, the expected utility interval $[\underline{\mathbb{E}}_m(f), \overline{\mathbb{E}}_m(f)]$ is reduced to a point, making the set-valued prediction a precise one. This is clearly an important difference as compared to the approaches described in Section 3.3, which can yield set-valued predictions with even precise probabilities of states of nature.

Example 3. *Consider again the binary classification problem in Example 2. Take* $m(\{\omega_1\}) = 0.5$, $m(\{\omega_2\}) = 0.3$, $m(\{\omega_1, \omega_2\}) = 0.2$ *as an example. Since only precise assignments are considered, we calculate the expected utility intervals* $[\underline{\mathbb{E}}_m(f_1), \overline{\mathbb{E}}_m(f_1)] = [0.60\ 0.76]$ *and* $[\underline{\mathbb{E}}_m(f_2), \overline{\mathbb{E}}_m(f_2)] = [0.51\ 0.65]$. *According to the strong dominance, maximality, e-admissibility and weak dominance criteria, we obtain, respectively, the choice sets* $\mathcal{F}^*_{SD} = \{f_1, f_2\}$, $\mathcal{F}^*_{max} = \{f_1, f_2\}$, $\mathcal{F}^*_{ead} = \{f_1, f_2\}$ *and* $\mathcal{F}^*_{WD} = \{f_1\}$. *The first three criteria provide a set-valued prediction. For the interval-valued utility criterion with* $\alpha_u = 0.7$ *and* $\beta_u = 0.3$, *we obtain the expected utility intervals* $[\underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_1), \overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_1)] = [0.648\ 0.712]$ *and* $[\underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_2), \overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_2)] = [0.552\ 0.608]$. *Therefore, this criterion induces a precise prediction* $\mathcal{F}^*_{IVU} = \{f_1\}$. *We can remark that the interval* $[\underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i), \overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i)]$ *is narrower than* $[\underline{\mathbb{E}}_m(f_i), \overline{\mathbb{E}}_m(f_i)]$, *as the latter corresponds to the bounds of interval obtained with* $\alpha_u = 1$ *and* $\beta_u = 0$.

When the mass function is Bayesian, such as $m(\{\omega_1\}) = 0.8$ *and* $m(\{\omega_2\}) = 0.2$, *there is no uncertainty in the states of nature. In this case,* $\underline{\mathbb{E}}_m(f_1) = \overline{\mathbb{E}}_m(f_1) = \underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_1) = \overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_1) = 0.84$, $\underline{\mathbb{E}}_m(f_2) = \overline{\mathbb{E}}_m(f_2) = \underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_2) = \overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_2) = 0.44$. *We then obtain a complete preorder* $f_1 \succ f_2$ *and a precise prediction* f_1 *with all the five criteria.*

[4] An *allocation* of a mass function m is defined as a mapping $a : \Omega \times (2^\Omega \setminus \{\emptyset\}) \to [0, 1]$, such that $\sum_{\omega \in K} a(\omega, K) = m(K)$, $\forall K \subseteq \Omega$.

3.5. Discussion on time complexity

In this section, we discuss the time complexity of the proposed approach. The discussion involves both generating the extended utility matrix $\widetilde{\mathbf{U}} = (\widetilde{u}_{K,j})_{(2^n-1)\times n}$ from $\mathbf{U} = (u_{ij})_{n\times n}$ and classifying a new instance.

The computation of the extended utility matrix needs to be performed only once for a given classification problem. To extend the utility matrix based on the OWA approach, we need to determine the OWA weights by optimization and then calculate each $\widetilde{u}_{K,j}$ by weighted sum operation (vector multiplication). The OWA-related nonlinear optimization problem can only be solved by an iterative algorithm, making its time complexity difficult to characterize. However, the determination of the OWA weights only depends on γ and the cardinality of set K. So we can compute weights in advance and tabulate for different values of γ (such as 0.5,0.55,0.6,..., 0.95,1) and different cardinalities (like 2,3,...). Therefore, the complexity of the optimization part does not really matter.

Given the OWA weights, basically, to calculate the aggregated utilities $\widetilde{u}_{K,j}$ for each nonempty subset K given each state of nature ω_j using Equation (11), we need to compute $n(2^n - 1 - n)$ weighted sums. However, we do not need to consider all subsets of Ω every time, as it will obviously become infeasible for large n. When n is large, we can consider only the pairs and Ω, which brings the number of subsets to $n(n-1)/2 + 1$. Considering each state of nature, we need to compute $n[n(n-1)/2 + 1]$ weighted sums in total, with a time cost of $O(n^3)$. It should be noticed that the base utility matrix $\mathbf{U}$ will be the identity matrix most of the time. In this case, the time complexity can be reduced to $O(n^2)$, as we only need to fill in non-zero elements $2 \times n(n-1)/2 + n$ times.

Considering the classification of each instance, the time complexity depends on the decision criterion. For partial classification via complete preorders among partial assignments, we need to compute a certain kind of expected utility for each of the $2^n - 1$ (for small n) or $n + \frac{n(n-1)}{2} + 1$ (for large n) acts and find the maximum one, which leads to a time complexity of $O(n^2)$. For partial classification via partial preorders among precise assignments, the E-admissibility criterion involves a linear programming for each act, which makes it much more time-consuming. Weak dominance and interval-valued utility criteria have a time complexity of $O(n)$. Strong dominance and maximality criteria compare the expected utility intervals for each pair of acts, with time complexity $O(n^2)$.

3.6. Evaluation of set-valued predictions

In classification applications, a test set $\mathcal{T}$ is typically used to assess the performance of a trained classifier. To choose a proper criterion, a standard is needed to evaluate different decisions. Traditional accuracy becomes improper when set-valued predictions are allowed.

Zaffalon [51] proposes a utility-discounted predictive accuracy under the $\{0, 1\}$ reward assumption to evaluate set-valued predictions made by credal classifiers. For a set-valued prediction $K \subseteq \Omega$ consisting of k classes, the *discounted accuracy* is defined as $\frac{1}{k} I(\omega \in K)$, where $I(\cdot)$ is the indicator function and ω its true label. A utility function then maps discounted accuracy to utility. For instance, the utility function u_{65} is defined as

$$u_{65}(x) = -0.6x^2 + 1.6x,$$

where x is the discounted accuracy of the issued prediction. This quadratic function is specified by three points: $u(1) = 1$ (the utility of a correct and precise prediction should be fixed at one), $u(0) = 0$ (the utility of wrong classification is zero), and $u(0.5) = 0.65$ (a certain utility of $x = 0.5$ is given to reveal the DM's attitude of risk aversion). The *utility-discounted accuracy*, which is biased considering personal preferences, is selected as a predictive accuracy in the test set:

$$\mathsf{Acc}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} u_{65} \left(\frac{1}{|K(i)|} I(\omega_{(i)} \in K(i)) \right),$$

where $K(i)$ is the predicted set of classes for test instance i, and $\omega(i)$ is its true class. This approach works well with the $\{0, 1\}$ reward case, but it cannot be generalized straightforwardly to more general cases.

In this paper, we consider the case of utilities in $[0, 1]$. We aim to evaluate precise or set-valued predictions by a single number (rather than a vector of parameters as proposed in [5]), with the requirement that the better the prediction, the larger the performance measure. In Section 3.1, we have seen how to extend a utility matrix so as to compute the utility of partial assignments. Accordingly, we propose to evaluate the classification performance by the *averaged utility* of the decisions made for the instances in test set $\mathcal{T}$:

$$\mathsf{AvU}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \widetilde{u}_{K(i),\omega(i)}. \tag{21}$$

The calculation of $\widetilde{u}_{K(i),\omega(i)}$ can be done using an OWA operator as proposed in Section 3.1, with an imprecision tolerance degree γ that models the DM's attitude.

4. Experiments and discussions

Classification experiments were carried out to compare the different decision strategies discussed in Section 3: precise classification with and without rejection, as well as partial classification based on complete or partial preorders. Section 4.1 first provides an intuitive comparison of decision criteria based on the UCI Iris data set. In Section 4.2, taking different values of γ to extend the utility matrix, we present and discuss the classification performances on UCI data sets. Considering simulated Gaussian data sets, the performances with noisy test sets are demonstrated in Section 4.3. Finally, Section 4.4 is dedicated to parameter selection in the generalized Hurwicz, OWA and interval-valued utility criteria.

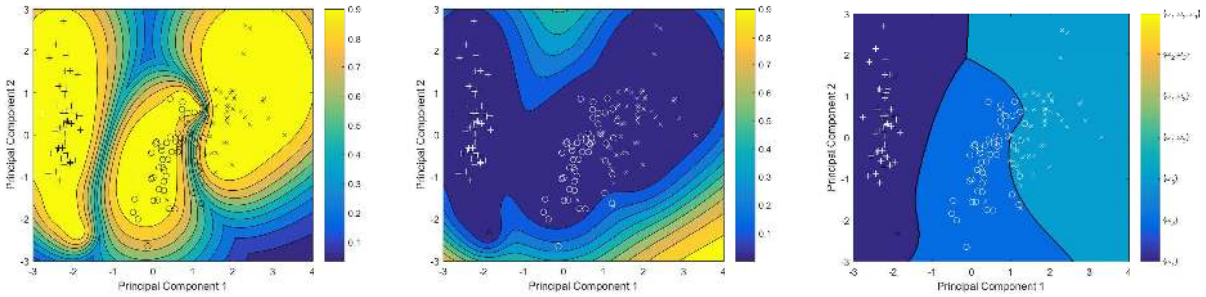
The mass functions concerning the states of nature were generated by the evidential neural network [10] recalled in Appendix A. The network was set with six prototypes per class, each prototype having full membership to only one class. The initial prototype locations were determined by the k-means clustering algorithm started from random initial conditions. We set the parameters with initial values $\alpha^i = 0.5$ and $\gamma^i = 0.1$ for $i = 1, \cdots, k$. A regularization parameter tuning the imprecision of the network output was fixed at $\rho = 0.99$ (with $m^i(\Omega) = 1 - \rho\alpha^i \exp(-\gamma^i (d^i)^2)$, $\forall i \in \{1, \cdots, k\}$). Note that the mass functions generated by this classifier have the property that their focal sets include only singletons and Ω, which

makes calculations easier. However, more general mass functions such as those generated by a multilayer perception [14] or the contextual discounted k-nearest neighbor rule [16] could be considered as well.

4.1. Iris data set

The UCI Iris data set was first taken as an illustrative example to visualize the behavior of different decision criteria. Considering that there are $n = 3$ classes, the utility matrix was assumed to be the 3×3 identity matrix. Mass functions were generated by the evidential neural network classifier. The mass function for each instance thus contained the masses given to each singleton and the frame of discernment, *i.e.*, $m(\{\omega_i\}) = m_i$, $i = 1, \cdots, n$ and $m(\Omega) = 1 - \sum_{i=1}^{n} m_i$ (see Figures 3a-3b).

Given this specific form of mass functions, all the three precise classification criteria discussed in Section 3.2 (Generalized maximin/maximax criteria and Pignistic criterion) result in the same hard partition of the instance space as shown in Figure 3c. The x-axis and y-axis correspond, respectively, to the first and second principal components of Iris instances. For any new pattern, its label is predicted based on the space partition. From Figure 3c, it can be seen that for some areas near the decision boundaries and areas far from the training instances, a precise prediction has a high probability of misclassification.

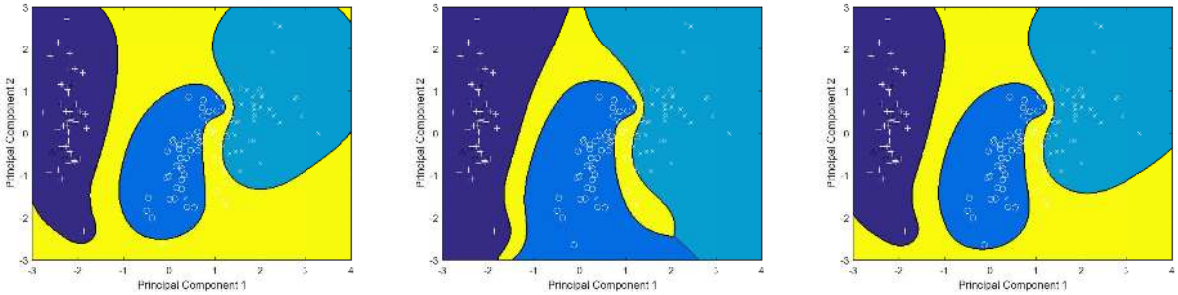


(a) $m(\{\omega_i\})$ (b) $m(\Omega)$ (c) Decision regions

Figure 3: Iris data set: contours of mass functions provided by the evidential neural networks (a,b) and decision regions for the decision rules with precise assignment (c).

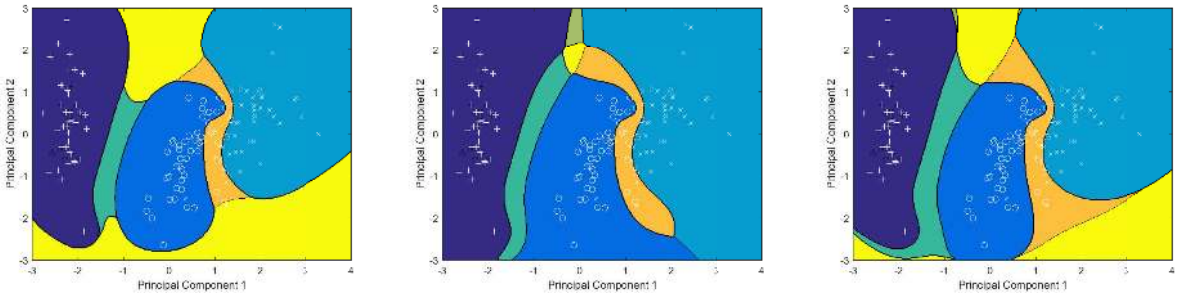
Figure 4 displays the decisions according to three criteria with rejection in Section 3.2 ($\lambda_0 = 0.8$). These figures share the same colorbar as Figure 3c, similarly for the Iris figures hereafter. Compared to Figure 3c, patterns situated close to the boundaries tend to be rejected, making it possible to achieve lower error rates. Moreover, the more cautious Maximin and Pignistic criteria reject patterns that are far away from any training instances.

Compared to mere rejection, partial classification can provide more informative results. Consider the criteria via complete preorders among partial assignments in Section 3.3. The utility matrix is extended via an OWA operator with $\gamma = 0.8$. Parameters of decision criteria are set to be $\alpha = 0.6$ (Hurwicz criterion) and $\beta = 0.6$ (OWA criterion). Figure 5 shows the decision regions induced by the different criteria.

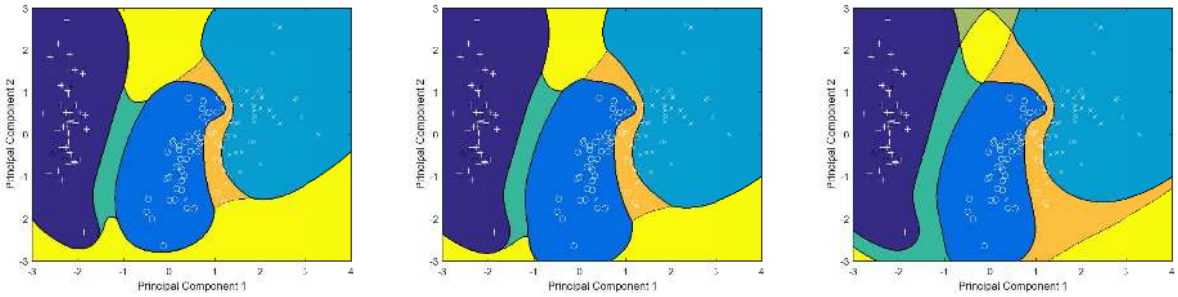


(a) Maximin (b) Maximax (c) Pignistic

Figure 4: Iris data set: decision regions with rejection (obtained for $\lambda_0 = 0.8$).



(a) Maximin (Hurwicz $\alpha = 1$) (b) Maximax (Hurwicz $\alpha = 0$) (c) Pignistic (OWA $\beta = 0.5$)



(d) Minimax regret (e) Hurwicz ($\alpha = 0.6$) (f) OWA ($\beta = 0.6$)

Figure 5: Iris data set: partial classification (complete preorder case) with the six decision criteria.

As compared to precise predictions without or with rejection, more reasonable decisions can be reached by partial classification. Generally speaking, partial classification can make a distinction between information conflict (appearing in areas near the decision boundaries) and lack of knowledge (areas far from the training instances). As shown in Figures 5a and 5d, the maximin and minimax regret criteria do give the same results, as Proposition 1 stated. Different from other criteria, the maximax criterion yields much more precise predictions (Figure 5b). Yet as compared to Figure 4b, partial classification allows us to provide specific information about the class label (say, class 2 or class 3), rather than plain rejection: more

informative predictions are made by our approach.

Figure 6 shows the decision boundaries for the Iris data set obtained with partial preorders among precise assignments (Sction 3.4). For the interval-valued utility criterion, we set $\alpha_u = 0.7$ and $\beta_u = 0.3$. For this data set, there is no difference among some criteria. The weak dominance and interval-valued utility criteria always give precise predictions, while the other three criteria can differentiate conflict from lack of information. Compared to previous results based on complete preorders of partial assignments (Figure 5), fewer set-valued predictions are made here, taking less advantage of uncertain information.

(a) Strong dominance, maximality, e-admissibility (b) Weak dominance, interval-valued utility

Figure 6: Iris data set: partial classification (partial preorder case) with the five decision criteria.

4.2. Classification performances with varying γ

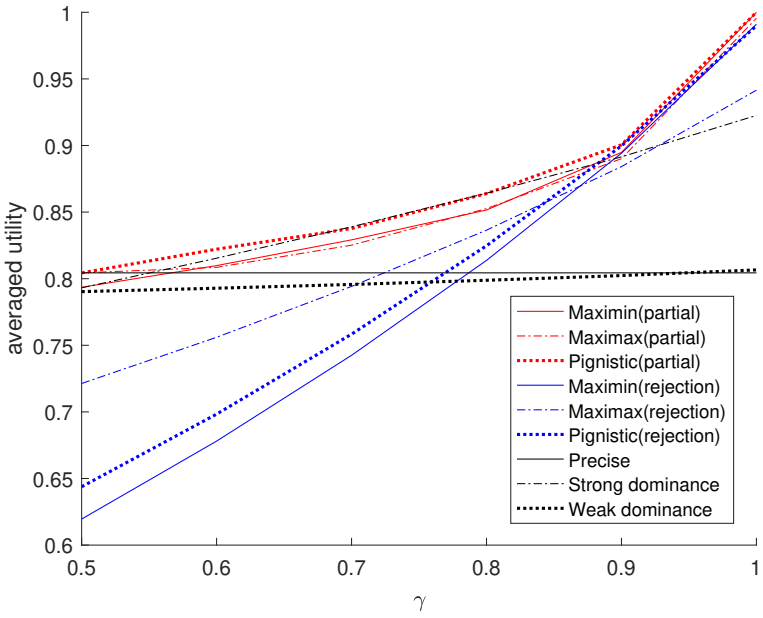
We then checked the averaged utilities obtained from utility matrices extended with varying imprecision tolerance degree γ. Experiments were carried out using 22 data sets from the UCI repository [25] with characteristics summarized in Table 3. The original utility matrix $\mathbf{U}$ was assumed to be the identity matrix of size n (the number of classes). The evidential neural network was set with six prototypes per class, expect for Lung cancer and Annealing data sets with three prototypes per class. Attributes or instances with more than 30% missing values were removed. To evaluate the performances, five-fold cross-validation was performed, and all experiments were repeated five times to compute an average result.

In addition to precise classification with and without rejection, several representative criteria were selected for partial classification. Considering complete preorders among partial assignments, we focussed on three decision criteria: Maximin (Hurwicz with $\alpha = 0$), Maximax (Hurwicz with $\alpha = 1$) and Pignistic (OWA with $\beta = 0.5$). Performances with other values of α and β are discussed in Section 4.4. The strong dominance, maximality and e-admissibility criteria yield similar results. Also, the interval-valued utility criterion performs similarly to weak dominance (as will be discussed in Section 4.4). Consequently, for partial classification via partial preorders among precise assignments, results obtained with only two criteria (strong and weak dominance) are reported.

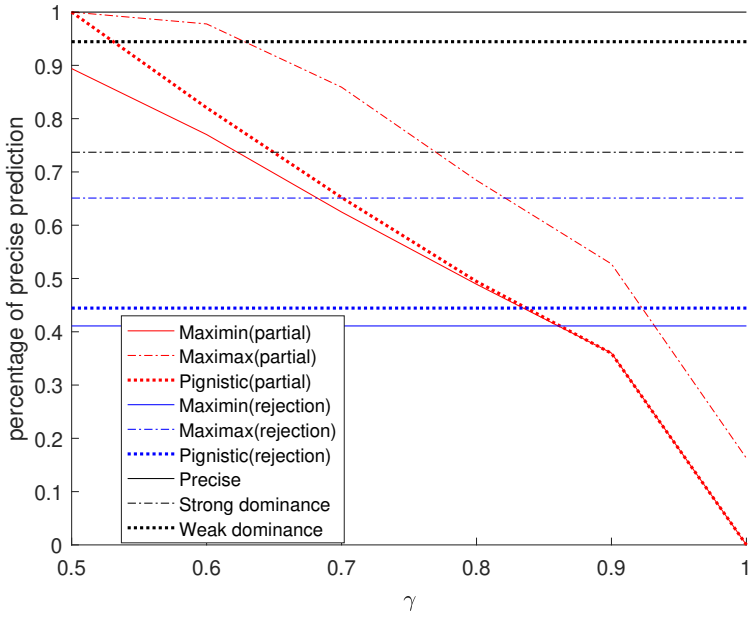
The averaged utilities and corresponding percentages of precise prediction for the Wine and Breast tissue data sets are shown, respectively, in Figures 7 and 8. In general, partial classification methods based on complete preorder (red curves) work better than those based

Table 3: UCI data sets for validation.

Data set	# of attributes	# of classes	# of instances
Adult	14	2	32526
Annealing	38	5	798
Balance Scale	4	3	625
Breast Tissue	9	6	106
Car Evaluation	6	4	1728
Contraceptive Method Choice	9	3	1473
Dermatology	34	6	358
Drug Consumption (Cannabis)	12	7	1885
Ecoli	8	7	336
Forest Mapping	27	4	523
Harberman's Survival	3	2	306
Hayes-Roth	4	3	160
Hepatitis	19	2	155
Image Segmentation	19	7	2310
Ionosphere	34	2	351
Iris	4	3	150
Lung Cancer	56	3	32
Mushroom	22	2	8124
SPECT Heart	22	2	267
Wine	13	3	178
Wine Quality (Red)	11	5	1599
Wireless Indoor Localization	7	4	2000

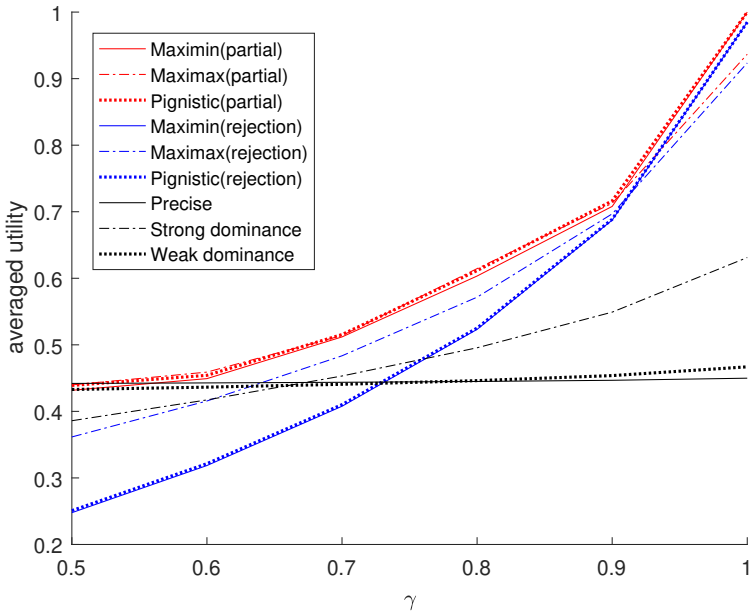


(a) Averaged utility

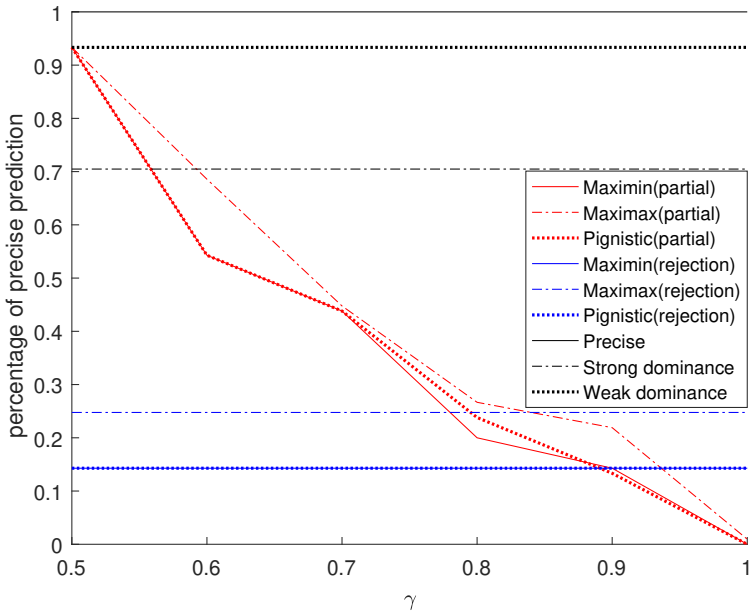


(b) Percentage of precise prediction

Figure 7: Experimental results with varying γ (Wine Data set).



(a) Averaged utility



(b) Percentage of precise prediction

Figure 8: Experimental results with varying γ (Breast tissue data set).

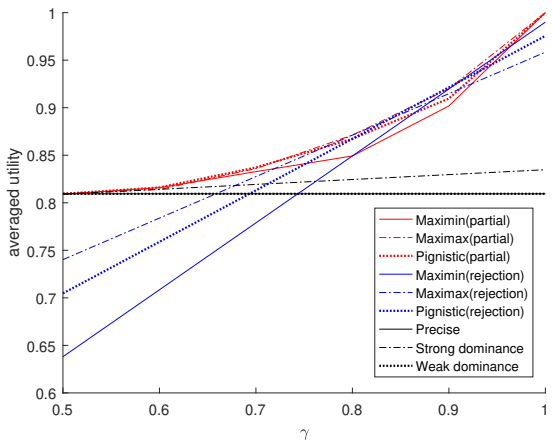
on partial preorder. They also perform better than rejection (blue curves). Compared to precise classifications, rejection methods have lower averaged utilities for small γ and higher ones for large γ. This general tendency can be explained as follows.

As we used the approach described in Section 3.1 to extend the utility matrix, the utility of a given set-valued prediction increases with γ. In our experiments, performance evaluation and partial classification via complete order of partial assignments shared the same γ. For full classification, neither the predicting procedure nor utilities of precise predictions are influenced by γ. For precise classification with rejection ($\lambda_0 = 0.8$), the vacuous predictions for some instances remain unchanged as γ varies from 0.5 to 1, but their utilities do grow from $1/n$ to 1, resulting in a higher averaged utility. A similar observation can be made for partial classification based on partial preorders, except that vacuous predictions are replaced by set-valued ones. It can be noted that the average utility of the weak dominance criterion increases mildly as γ grows, since it makes precise predictions most of the time.

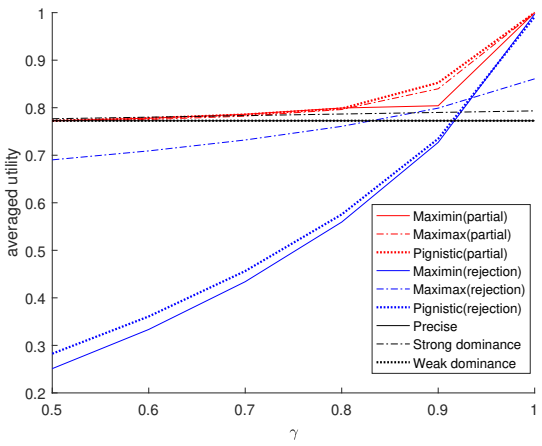
For partial classification based on complete preorders among partial assignments, given a fixed γ, the averaged utilities of different criteria vary in a small range, as shown by the red lines in Figures 7a and 8a. Among the Maximin, Maximax and Pignistic criteria, we can hardly conclude generally which criterion is the best. When γ increases, decisions change as set-valued predictions are more preferred. Averaged utilities increase monotonically with γ. We can also comparing a partial classification criterion with its corresponding one with rejection (such as the red and blue dotted lines in Figure 7a). It can be known from the $\widetilde{\mathbf{U}}$ that the more imprecise is the prediction, the lower utility it achieves. When γ is relatively small, say $\gamma \leq 0.8$, partial classification outperforms rejection, as it provides imprecise predictions rather than vacuous ones. When γ is close to 1, partial classification makes much more imprecise predictions. Together with the fact that utility of vacuous prediction increases to 1, partial classification achieves an averaged utility approaching 1.

For the other data sets, as shown in Figures 9-12, partial classification with complete preorders generally outperforms the rejection approach (although, for some sets such as the Adult (Figure 9a) and Balance scale (Figure 9c) data sets, this is only true for small γ). The strong dominance criterion works the best for Dermatology (Figure 9f) and Wireless indoor localization (Figure 12b) data sets. The weak dominance criterion, which usually produces precise predictions, yields quite similar result as precise classification for most data sets. Strong dominance may outperform weak dominance for all γ (as with the Adult (Figure 9a) and Car evaluation (Figure 9d) data sets), or only for large γ (as with the Drug consumption (Figure 10a) and Hayes-roth (Figure 10e) data sets). For some datasets such as Balance scale (Figure 9c) and Iris (Figure 11c), strong dominance and weak dominance can also have the same performance for any γ. Overall, the main finding is that partial classification with complete preorders (with the maximin, maximax of pignistic criterion) outperform the other criteria for most of the datasets.

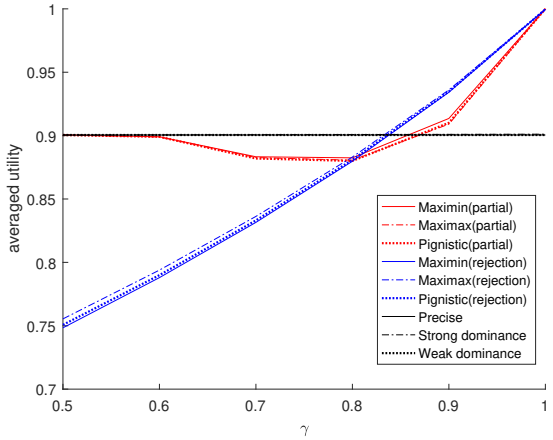
There is something special for some data sets such as Balance scale, Dermatology and Iris: Considering partial classification based on complete preorders (red curves in Figures 9c, 9f and 11c), the averaged utility curves have a U-shape (they first decrease when γ increases, before increasing again). Consider the Iris data set as an example. Here in Figure 11c, only the given 150 instances marked with $+$, $\times$ and $\circ$ in Figure 3c are used for training and test.



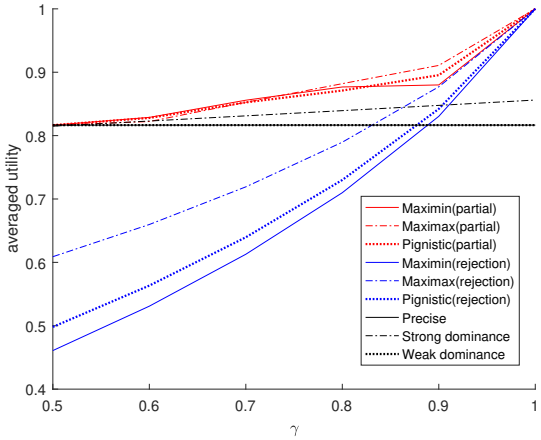
(a) Adult data set



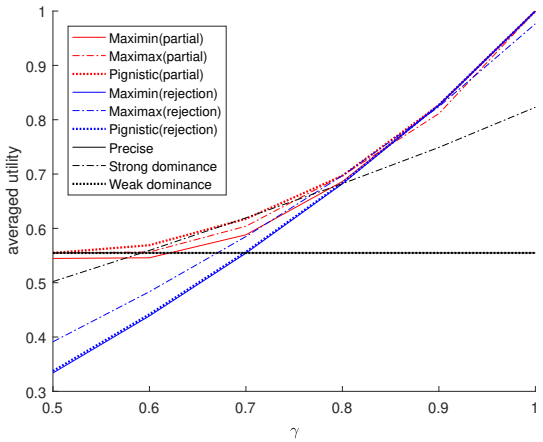
(b) Annealing data set



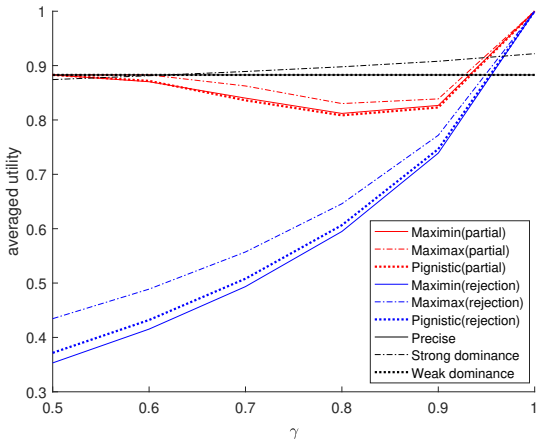
(c) Balance scale data set



(d) Car evaluation data set

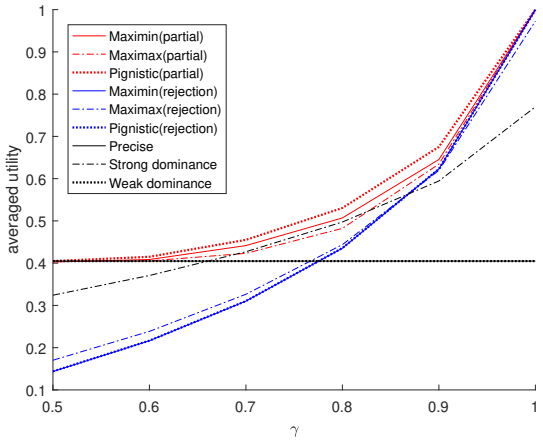


(e) Contraceptive method choice data set

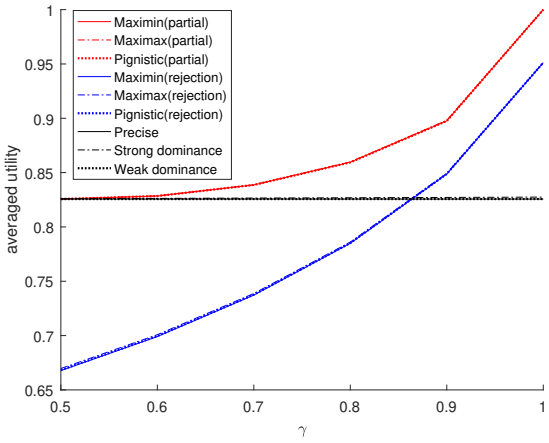


(f) Dermatology data set

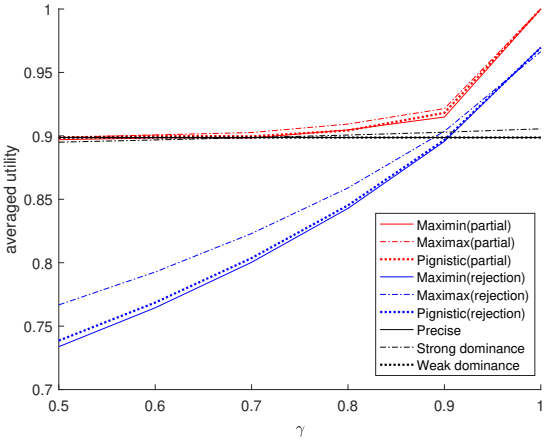
Figure 9: Averaged utilities with varying γ (part 1).



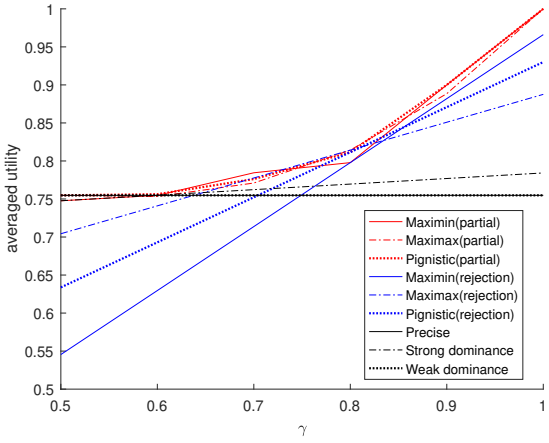
(a) Drug consumption data set



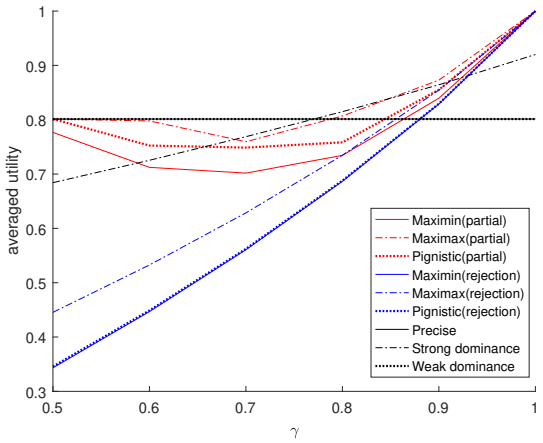
(b) Ecoli data set



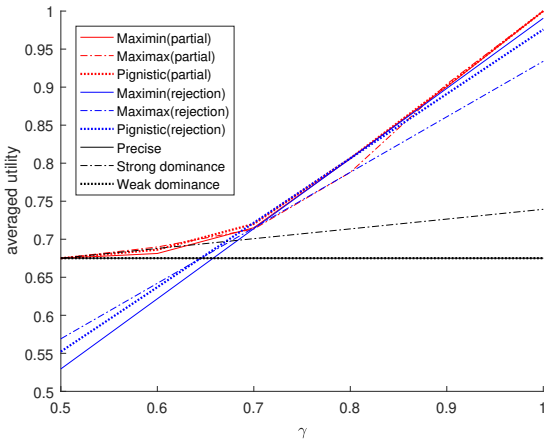
(c) Forest mapping data set



(d) Harberman's survival data set

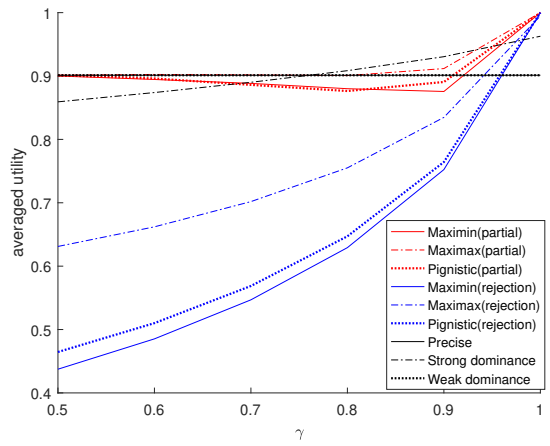


(e) Hayes-roth data set

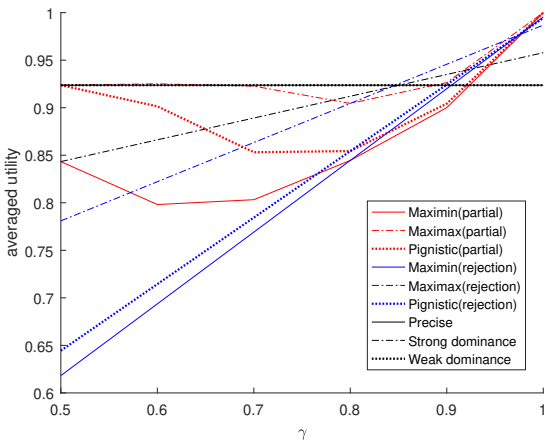


(f) Hepatitis data set

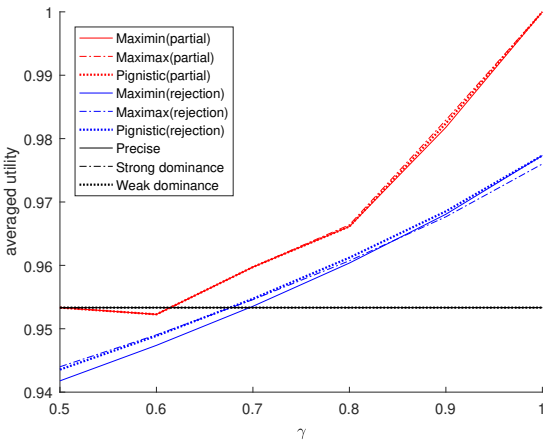
Figure 10: Averaged utilities with varying γ (part 2).



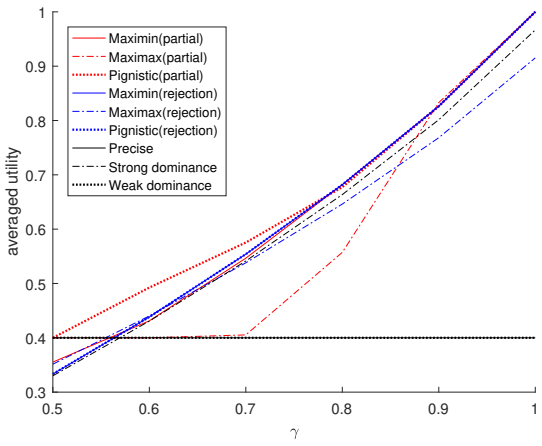
(a) Image segmentation data set



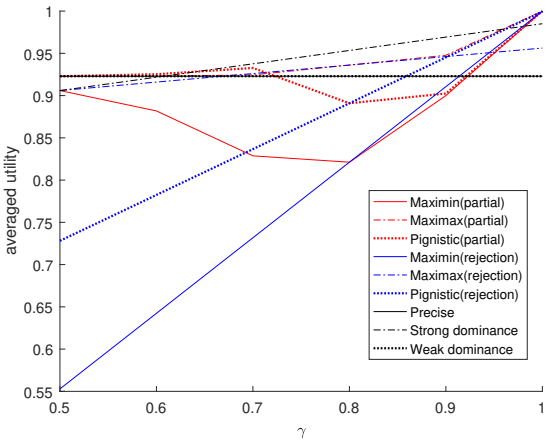
(b) Ionosphere data set



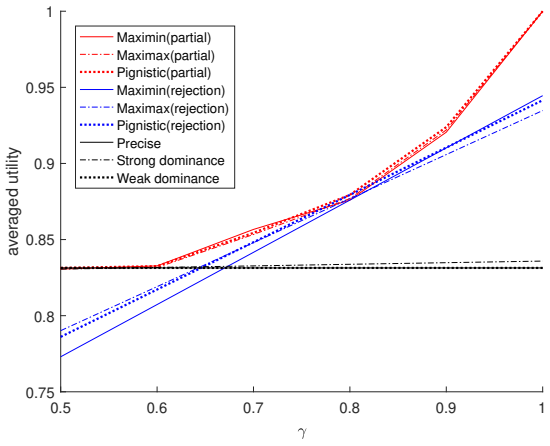
(c) Iris data set



(d) Lung cancer data set

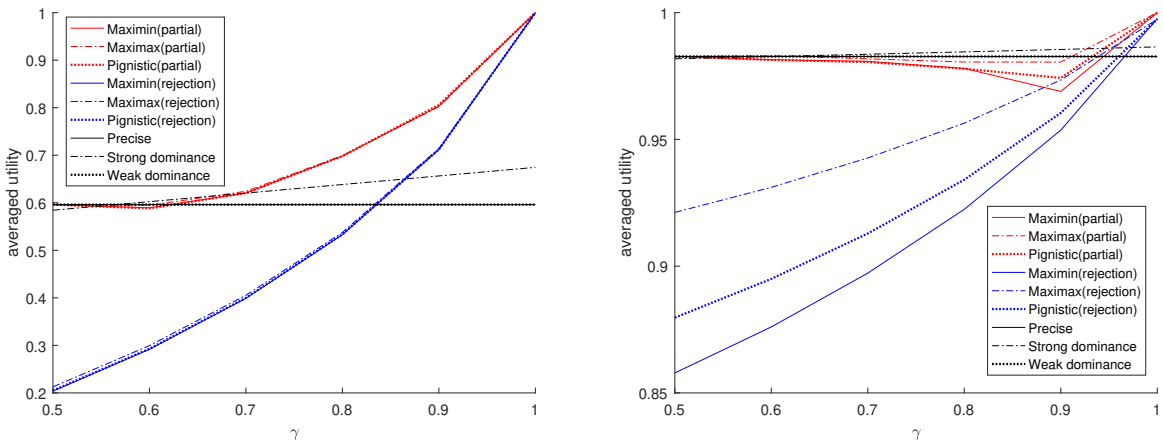


(e) Mushroom data set



(f) SPECT heart data set

Figure 11: Averaged utilities with varying γ (part 3).



(a) Wine quality data set (normalized)

(b) Wireless indoor localization data set

Figure 12: Averaged utilities with varying γ (part 4).

Although partial classification does induce a better partition of the input space, it has little effect on the 150 training instances which are well separated in different classes. In such a case, hard partitioning of the input space guarantees a good classification performance. When we perform partial classification with complete preorders among partial assignments, we obtain some correct but imprecise predictions, which have lower utilities than the precise and correct ones. In general, partial classification is more beneficial when different classes overlap in the selected training set. Taking the Breast tissue data set as an illustration (Figure 8), partial classification yields a particular advantage over other decision methods due to their cautiousness. They provide larger sets (but not vacuous ones) including the true label ω^* rather than smaller sets excluding ω^*.

4.3. Performances with Noisy Test Sets

In many situations, a classifier is trained with "good" data (acquired and preprocessed in controlled conditions) and then used in a real environment where, for instance, sensors may not be well calibrated. In such a case, the test data do not have the same distribution as the learning data. Cautious decision rules making set-valued predictions can be expected to be particular beneficial in such an environment, and discrepancies between the performances of different decisions rules may be more apparent than they are in the case of "clean" data considered in previous experiments.

To validate this hypothesis, we performed the experiments on an artificial Gaussian data set. Considering a three-class problem with data set of two attributes, the training set was simulated from three Gaussian distributions with the following characteristics:

$$\mu_1 = [-1, 0]^T, \mu_2 = [1, 0]^T, \mu_3 = [2, 1]^T,$$

$$\sigma_1 = 0.25I, \sigma_2 = 0.75I, \sigma_3 = 0.5I,$$

where I is the identity matrix. Figure 13 visualizes a particular data set of 600 instances generated in this way.

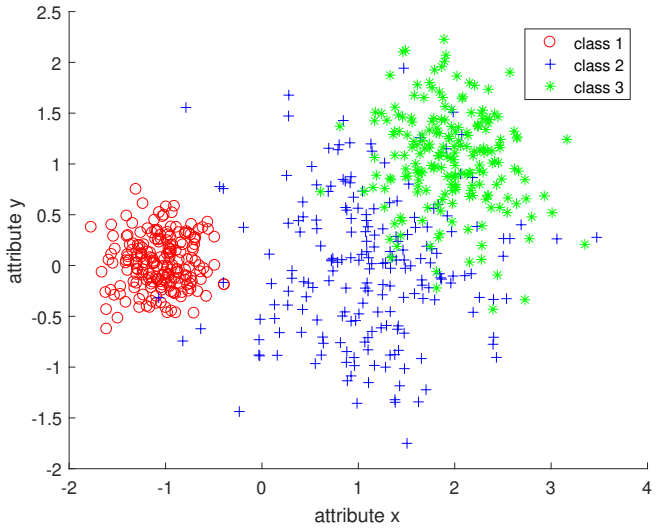


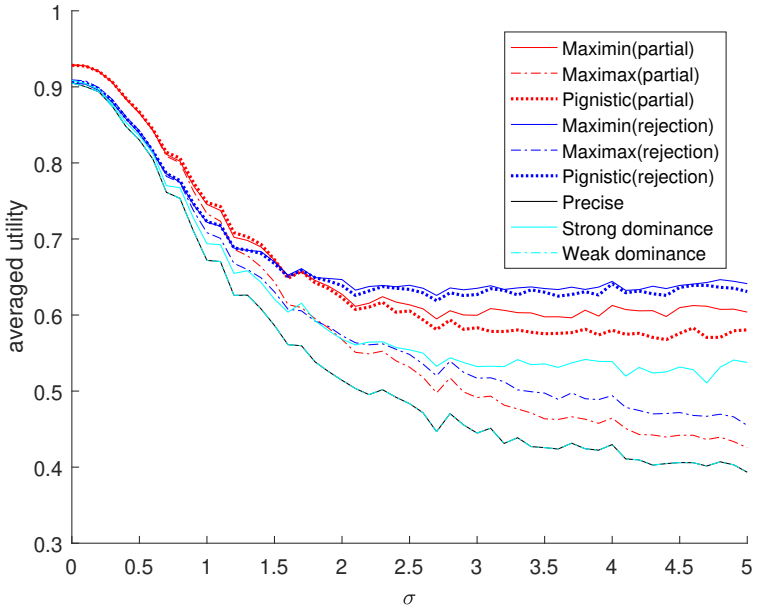
Figure 13: A Gaussian data set of 600 instances.

Algorithm 1: Algorithm to generate a noisy test set.

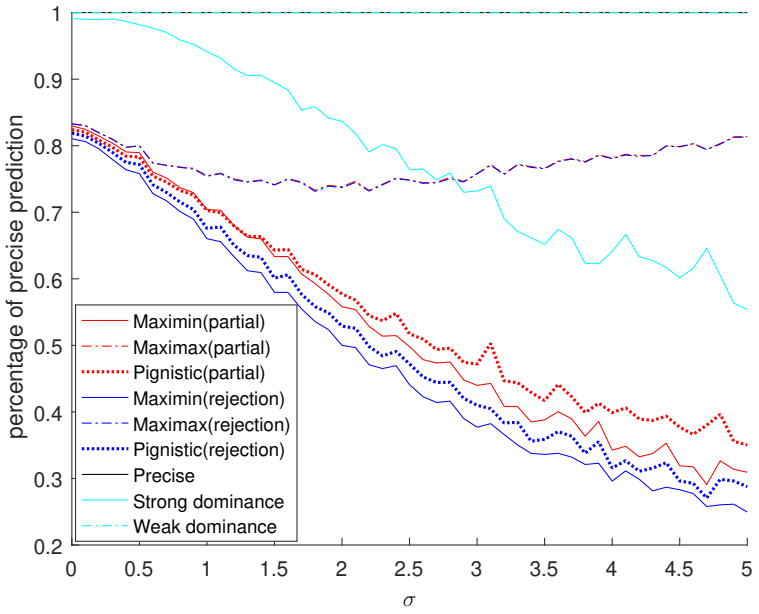
Input: test set $T = \{(\boldsymbol{x}, \boldsymbol{y}), C\}$, noise standard deviation σ
Output: noisy test set $\tilde{T} = \{(\tilde{\boldsymbol{x}}, \boldsymbol{y}), C\}$
for $1 \leq i \leq |T|$ **do**
    Draw $\epsilon(i)$ from $\mathcal{N}(0, \sigma^2)$;
    $\tilde{x}(i) = x(i) + \epsilon(i)$;

To simulate a different distribution, random noise was added to the features of the test instances using Algorithm 1. We set $\gamma = 0.8$ and let the noise standard deviation σ vary from 0 to 5 to simulate different levels of noise. The experiments were repeated 20 times to compute an average result. In each experiment, the training and test sets contain, respectively, 600 and 300 instances. With higher noise level, the distribution of the test set becomes more different from that of the training set. The averaged utilities are plotted against the noise level according to various decision criteria in Figure 14a. Similar to the previous experiments, the percentage of precise predictions shown in Figure 14b helps to analyze the performances.

For $\sigma = 0$, the test and training sets have the same distribution; partial classification via complete preorders among partial assignments (Section 3.3) outperforms other partial classification approaches (Section 3.4) as well as precise classifications with and without rejection slightly. When σ increases, the averaged utilities for all criteria drop quickly and the performances of different criteria in each approach start to differ. As the test



(a) Averaged utility



(b) Percentage of precise predictions

Figure 14: Results of different criteria as a function of noise level.

set distribution becomes more different, the precise approach yields the worst performance. Weak dominance makes precise predictions in almost all the cases, performing quite similarly to the precise approach.

Basically, when uncertainty increases, the decision criteria (except Maximax) assign more instances to sets. As different classes overlap more with larger σ, imprecise predictions are more likely to contain the true labels, achieving a higher averaged utility. The Maximax in both settings and strong dominance make more precise predictions than others, leading to a worse performance (but still better than the precise approach and weak dominance). For Maximin and Pignistic in both partial classification and rejection settings, when σ grows from 2.5 to 5, the averaged utilities remain steady or even increase slightly. The Maximin criterion is the most conservative one in both settings, resulting in the highest averaged utility.

Comparing the two approaches of partial classification, generally the family considering complete preorders among partial assignments is more suitable for noisy environments. We can further compare the complete preorder-based approach with the rejection strategy (the Maximin and Pignistic). The partial criteria always make less imprecise predictions than the rejection approaches. When σ is small, partial classification performs better as it provides correct and more accurate predictions; When σ is large, rejection is a better choice since vacuous prediction yields higher utility than wrong set-valued predictions.

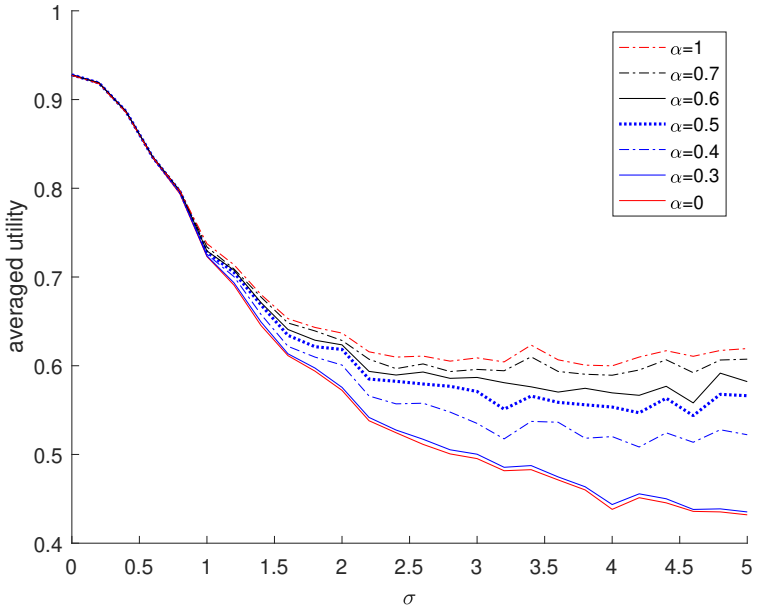
4.4. Parameter selection of decision criteria

In the fourth experiment, we study the problem of parameter selection in several decision criteria. Consider again the simulated Gaussian data set, we kept the same data setting as in Section 4.3 and we set $\gamma = 0.8$. Figures 15a and 15b show, respectively, the decisions made by the generalized Hurwicz and OWA criteria with different values of α and β.

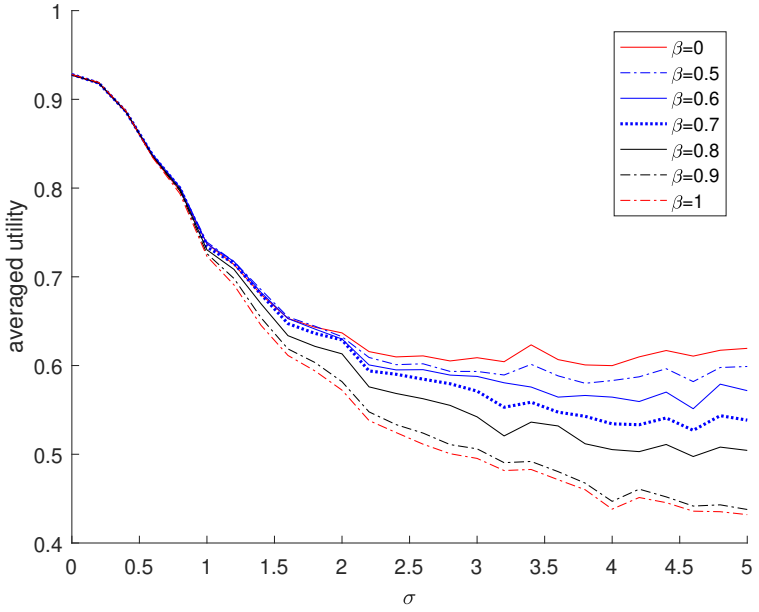
When the test and training sets have similar distributions, the value of the parameter in the two decision criteria has little effect on the classification performance. Differences become obvious as σ grows. We mainly analyze the case of high uncertainty (say, $\sigma \geq 2.5$). For the Hurwicz criterion, given a fixed σ, the averaged utility increases monotonically as α grows (as a reminder, we have Maximin when $\alpha = 1$ and Maximax when $\alpha = 0$). When $\alpha \leq 0.3$, most predictions are precise, leading to massive misclassifications. When $\alpha \geq 0.7$, the variation of α does not change averaged utility much.

For the OWA decision criterion, the averaged utility decreases monotonically as β grows. When $\beta = 0.5$ (pignistic criterion), the performance is quite acceptable. The further decrease of β leads to little improvement of performance. On the other side, the averaged utility drops quickly with $\beta > 0.8$, as the partitions of instance space tend to be hard ones. It can also be noted that both $\alpha = 1$ ($\alpha = 0$) for the Hurwicz criterion and $\beta = 0$ ($\beta = 1$) for the OWA criterion correspond exactly the Maximin (Maximax) criterion. By choosing α or β, performances can fall in between those of the Maximax and Maximin criteria.

We also investigated the interval-valued utility criterion by varying the two indices α_u and β_u. Consider the Gaussian data set with $\sigma = 0$, the experiments were repeated 20 times for an average result. Recall the preference relation that $f_i \succcurlyeq_{IVU} f_j \iff \Big(\underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i) \geq$



(a) Hurwicz criterion



(b) OWA criterion

Figure 15: Averaged utilities with changing criterion parameter for the Hurwicz (a) and OWA (b) criteria.

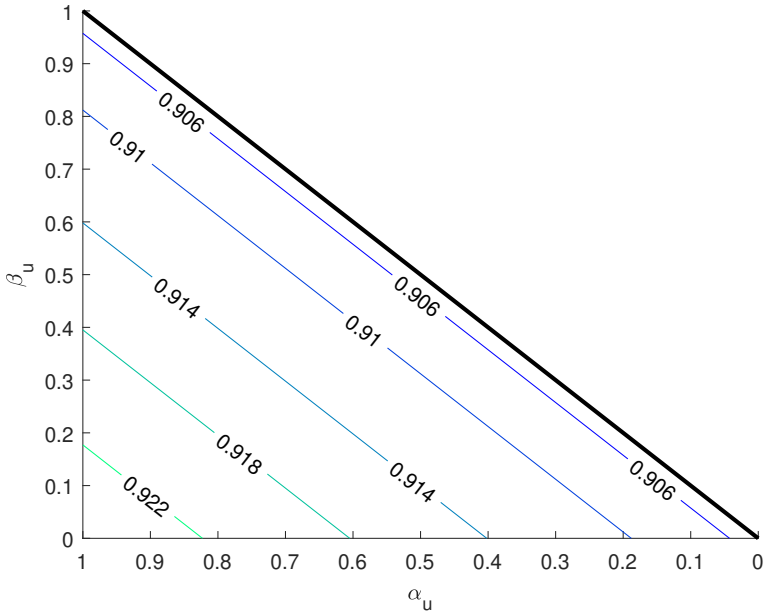


Figure 16: Averaged utilities with changing α_u and β_u.

$\underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_j)\Big)$ and $\Big(\overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i) \geq \overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_j)\Big)$, the underlying weak dominance leads to the averaged utility of 0.9052 for any $0 \leq \beta_u \leq \alpha_u \leq 1$. However, if we consider the strong dominance and let $f_i \succcurlyeq_{IVU} f_j \iff \underline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_i) \geq \overline{\mathbb{E}}_{m,\alpha_u,\beta_u}(f_j)$, the resulted averaged utilities can vary between 0.9052 and 0.9244 with different α_u and β_u, as shown in Figure 16. When $\alpha_u = \beta_u$, the utility interval is reduced to a point, making the averaged utilities equal to that of precise prediction without rejection. As the interval $[\beta_u, \alpha_u]$ becomes wider, more set-valued predictions will be made, yielding a higher averaged utility.

5. Conclusion

Modelling uncertainty in the belief function framework, we have carried out a thorough analysis of various decision-making criteria for evidential partial classification. By allowing for imprecision in certain regions of the input space, partial classification has been shown to yield a higher averaged utility as compared to precise classification with and without rejection. The extended utility matrix is instrumental in obtaining the complete preorder of partial assignments in decision-making, as well as evaluating the performances of different decision criteria. Therefore, we have proposed to generate utilities of partial assignments via an OWA operator, with one parameter γ controlling the DM's attitude towards imprecision. Partial classification can be implemented via complete preorders among partial assignments or via partial preorders among complete assignments. Both approaches have been analyzed theoretically and experimentally. Based on 22 UCI and simulated Gaussian data sets, experimental results suggest some guidelines for criteria choosing in classification problems: In general, partial classification via complete preorder among partial assignments achieves

a more reasonable partition of the instance space, leading to better performances. With regard to partial classification, the best decision criterion has to be decided depending on the data set. In noisy environments, more cautious rules should be preferred, such as the Maximin criterion among partial classification methods. In case of a very noisy test set, classification with rejection works best as the most conservative approach.

It should be noted that partial classification ideas in the Dempster-Shafer setting can also be carried out similarly in other settings such as the imprecise probability framework [3]. In future work, we will also consider Shafer's constructive decision theory [38], which does not rely on utility, and study its performance in classification tasks.

Acknowledgement

This research was supported by the China Scholarship Council and Shandong Provincial Natural Science Foundation ZR2018PF009. It was also supported by the Labex MS2T funded by the French Government through the program "Investments for the future" by the National Agency for Research (reference ANR-11-IDEX-0004-02).

Appendix A. Evidential neural network classifier

The *evidential neural network classifier* [10] is an adaptive classifier based on DS theory[5]. The classification procedure can be implemented in a specific neural network architecture with one input layer, two hidden layers and one output layer. Assessing the similarity of a pattern with a limited number of prototypes, items of evidence regarding the class membership are represented by mass functions and combined by Dempster's rule (4).

The evidential neural network classifier is similar to the evidential k nearest neighbor rule [8][16] but k prototypes $\boldsymbol{p}^1, \cdots, \boldsymbol{p}^k$ are considered instead of the nearest neighbors of a pattern $\boldsymbol{x}$ to reduce the computational complexity. Each prototype $\boldsymbol{p}^i$ provides a mass function m^i based on the Euclidean distance $d^i = \|\boldsymbol{x} - \boldsymbol{p}^i\|$ between $\boldsymbol{x}$ and $\boldsymbol{p}^i$, $i = 1, \cdots, k$. The unit mass is distributed among the singleton $\{\omega_q\}$ and Ω:

$$m^i(\{\omega_q\}) = \alpha^i u_q^i \exp(-\gamma^i (d^i)^2), \quad q = 1, \ldots, n \tag{A.1a}$$

$$m^i(\Omega) = 1 - \alpha^i \exp(-\gamma^i (d^i)^2), \tag{A.1b}$$

where γ^i is a scale parameter for prototype $\boldsymbol{p}^i$, u_q^i is the membership degree of prototype $\boldsymbol{p}^i$ to class ω_q (with $\sum_{q=1}^n u_q^i = 1$), and α^i is a parameter indicating the relative importance of prototype $\boldsymbol{p}^i$ in classifying new patterns. Combining the k mass functions $m^i, i = 1, \cdots, k$ by Dempster's rule (4), the resulted mass function describes the uncertainty pertaining to the class of a pattern.

Parameters in the model are trained by minimizing the mean squared differences between model outputs and target values. Once the neural network has been trained, an output

[5]This method is implemented in the R package `evclass` [12] available at `https://cran.r-project.org`. Matlab code can also be downloaded from `https://www.hds.utc.fr/~tdenoeux/dokuwiki/en/software/belief_nn`.

mass function can be computed for each test pattern, which conveys more information than does a probability distribution. Based on this mass function, we can further implement various decision strategies to make more reasonable predictions, such as classification with rejection [10] and partial classification discussed in Section 3.

[1] K. Ali, S. Manganaris, and R. Srikant. Partial classification using association rules. In *KDD'97: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, volume 97, pages 115–118, 1997.

[2] A. Appriou. Uncertain data aggregation in classification and tracking processes. In B. Bouchon-Meunier, editor, *Aggregation and Fusion of imperfect information*, pages 231–260. Physica-Verlag, Heidelberg, 1998.

[3] T. Augustin, F. P. Coolen, G. De Cooman, and M. C. Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[4] C.-K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957.

[5] G. Corani and M. Zaffalon. Lazy naive credal classifier. In *Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*, pages 30–37. ACM, 2009.

[6] J. J. Del Coz, J. Díez, and A. Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10(79):2273–2293, 2009.

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

[8] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[9] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern recognition*, 30(7):1095–1107, 1997.

[10] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.

[11] T. Denoeux. 40 years of Dempster-Shafer theory. *International Journal of Approximate Reasoning*, 79:1–6, 2016.

[12] T. Denoeux. *evclass: Evidential Distance-Based Classification*, 2017. R package version 1.1.1.

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

[14] T. Denœux. Logistic regression, neural networks and Dempster-Shafer theory: A new perspective. *Knowledge-Based Systems*, 176:54–67, 2019.

[15] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Beyond probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, volume 1, chapter 4, pages 119–150. Springer Verlag, 2020.

[16] T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta. A new evidential k-nearest neighbor rule based on contextual discounting with partially supervised learning. *International Journal of Approximate Reasoning*, 113:287–302, 2019.

[17] T. Denœux and P. P. Shenoy. An interval-valued utility theory for decision making with Dempster-Shafer belief functions. *International Journal of Approximate Reasoning*, 124:194–216, 2020.

[18] D. Filev and R. R. Yager. Analytic properties of maximum entropy OWA operators. *Information Sciences*, 85(1):11–27, 1995.

[19] E. Frank and M. Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156. Springer, 2001.

[20] G. Fumera, F. Roli, and G. Giacinto. Reject option with multiple thresholds. *Pattern recognition*, 33(12):2099–2101, 2000.

[21] T. M. Ha. The optimum class-selective rejection rule. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):608–615, 1997.

[22] R. Herbei and M. H. Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, 34(4):709–721, 2006.
[23] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2017.
[24] I. Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press, 1983.
[25] M. Lichman. UCI machine learning repository, 2013.
[26] X. Liu and L. Chen. On the properties of parametric geometric OWA operator. *International Journal of Approximate Reasoning*, 35(2):163–178, 2004.
[27] Z.-G. Liu, L. Huang, K. Zhou, and T. Denœux. Combination of transferable classification with multi-source domain adaptation based on evidential reasoning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2020.
[28] Z.-G. Liu, Q. Pan, J. Dezert, and A. Martin. Combination of classifiers with optimal weight based on evidential reasoning. *IEEE Transactions on Fuzzy Systems*, 26(3):1217–1230, 2018.
[29] Z.-G. Liu, Q. Pan, G. Mercier, and J. Dezert. A new incomplete pattern classification method based on evidential reasoning. *IEEE Transactions on Cybernetics*, 45(4):635–646, 2015.
[30] L. Ma and T. Denœux. Making set-valued predictions in evidential classification: A comparison of different approaches. In *International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA 2019)*, pages 276–285, Ghent, Belgium, July 2019. Published in Proceedings of Machine Learning Research 103:276–285, 2019.
[31] T. Mortier, M. Wydmuch, K. Dembczyński, E. Hüllermeier, and W. Waegeman. Efficient set-valued prediction in multi-class classification. *arXiv preprint arXiv:1906.08129*, 2019.
[32] M. O'Hagan. Aggregating template or rule antecedents in real-time expert systems with fuzzy set logic. In *Twenty-Second Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 681–689, 1988.
[33] E. Ramasso and R. Gouriveau. Prognostics in switching systems: Evidential markovian classification of real-time neuro-fuzzy predictions. In *2010 Prognostics and System Health Management Conference*, pages 1–10. IEEE, 2010.
[34] A. P. Reynolds and B. De la Iglesia. A multi-objective grasp for partial classification. *Soft Computing*, 13(3):227–243, 2009.
[35] M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.
[36] L. J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67, 1951.
[37] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, 1976.
[38] G. Shafer. Constructive decision theory. *International Journal of Approximate Reasoning*, 79:45–62, 2016.
[39] P. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005.
[40] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66(2):191–243, 1994.
[41] T. M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4(5–6):391–417, 1990.
[42] Z.-G. Su, T. Denœux, Y.-S. Hao, and M. Zhao. Evidential K-NN classification with enhanced performance via optimizing a class of parametric conjunctive t-rules. *Knowledge-Based Systems*, 142:7–16, 2018.
[43] M. C. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.
[44] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.
[45] P. Walley. *Statistical reasoning with imprecise probabilities*. Monographs on statistics and applied

probability. Chapman and Hall, 1991.
[46] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190, 1988.
[47] R. R. Yager. Decision making under Dempster-Shafer uncertainties. *International Journal of General Systems*, 20(3):233–245, 1992.
[48] R. R. Yager. Decision making using minimization of regret. *International Journal of Approximate Reasoning*, 36(2):109–128, 2004.
[49] G. Yang, S. Destercke, and M.-H. Masson. The costs of indeterminacy: How to determine them? *IEEE Transactions on Cybernetics*, 47(12):4316–4327, 2017.
[50] M. Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1):5–21, 2002.
[51] M. Zaffalon, G. Corani, and D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.