



**HAL**  
open science

# Partial Classification in the Belief Function Framework

Liyao Ma, Thierry Denoeux

► **To cite this version:**

Liyao Ma, Thierry Denoeux. Partial Classification in the Belief Function Framework. Knowledge-Based Systems, 2021, 214, pp.106742. 10.1016/j.knosys.2021.106742 . hal-03132379

**HAL Id: hal-03132379**

**<https://hal.science/hal-03132379>**

Submitted on 5 Feb 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Partial Classification in the Belief Function Framework

Liyao Ma<sup>a</sup>, Thierry Denœux<sup>b,c,d</sup>

<sup>a</sup>University of Jinan, Jinan, China

<sup>b</sup>Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France

<sup>c</sup>Shanghai University, UTSEUS, Shanghai, China

<sup>d</sup>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](mailto:cse_maly@ujn.edu.cn) (Liyao Ma), [thierry.denoeux@utc.fr](mailto:thierry.denoeux@utc.fr) (Thierry Denœux)

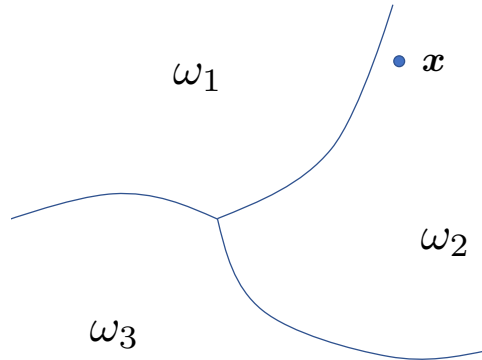


Figure 1: A situation of classification uncertainty with three decision regions. Pattern  $\mathbf{x}$  cannot be reliably assigned to classes  $\omega_1$  or  $\omega_2$ , but it almost certainly does not belong to  $\omega_3$ . The set-valued prediction  $\{\omega_1, \omega_2\}$  is a reliable option in this case.

13 too high, making it possible to reduce the probability of misclassification. However, there  
 14 are cases where, although the risk of misclassification is high, a *subset* of labels can still be  
 15 considered as very plausible. For instance, in the example illustrated in Figure 1, pattern  
 16  $\mathbf{x}$  cannot be classified with high certainty, but it almost certainly does not belong to class  
 17  $\omega_3$ . In such a case, assigning  $\mathbf{x}$  to the set of classes  $\{\omega_1, \omega_2\}$  seems to be a more reasonable  
 18 option than outright rejection. Such *partial*<sup>1</sup>, or *set-valued classification* makes it possible  
 19 to better reflect uncertainty and increase the reliability of classifiers. Classification with a  
 20 reject option can be seen as a special case of partial classification, rejection being equivalent  
 21 to assigning the pattern to the entire set of possible labels. Compared to rejection, partial  
 22 classification makes it possible to provide more informative decisions while still minimizing  
 23 the risk of misclassification.

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

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

<sup>1</sup>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.

<sup>2</sup>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.

35 classification is 1 and the utility of misclassification is 0, what is the utility of predicting the  
 36 set  $\{\omega_1, \omega_2\}$  if the true class is, say,  $\omega_1$ ? It should be strictly greater than 0, as the classifier  
 37 correctly eliminated class  $\omega_3$ , but strictly less than 1, as a correct but imprecise prediction is  
 38 arguably less valuable than a prediction that is both correct and precise. Here, we propose  
 39 a method to define the utilities of set-valued predictions based on the utility of precise ones  
 40 and a single control parameter, using Ordered Weighted Average (OWA) operators [46].  
 41 Having defined an extended utility matrix, a generalized Maximum Expected Utility (MEU)  
 42 principle allows us to make set-valued predictions based on different notions of expectation  
 43 with respect to a belief function. In the following, we briefly review previous approaches to  
 44 partial classification.

#### 45 *Related work*

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

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

67 In the imprecise probability framework, Zaffalon [50] introduces the naive credal classifier,  
 68 in which prior ignorance about the distribution of classes is modeled by means of a set of prior  
 69 densities (also called the prior *credal set*). This credal set is turned into a set of posterior  
 70 probabilities by element-wise application of Bayes’ rule. The classifier returns all the classes  
 71 that are non-dominated by any other class according to the posterior credal set. As we will  
 72 see in Section 3.4, a similar approach can be implemented with evidential classifiers. Zaffalon  
 73 et al. [51] propose a metric to evaluate the predictions of credal classifiers, considering the  
 74  $\{0, 1\}$  reward case and taking the decision maker’s degree of risk aversion into account.

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

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

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

## 102 2. Background

103 In this section, we review background notions and define the notations. The theory of  
104 belief functions is first reviewed in Section 2.1 and the OWA operators are recalled in Section  
105 2.2. The decision framework is defined in Section 2.3.

### 106 2.1. Theory of belief functions

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

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

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

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

122 The *belief* and *plausibility* functions corresponding to mass function  $m$  are defined, re-  
 123 spectively, as

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

124 and

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

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

130 Two mass functions  $m_1$  and  $m_2$  representing independent items of evidence can be com-  
 131 bined using *Dempster's rule* [37] as follows,

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

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

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

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

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

## 137 2.2. Ordered Weighted Average operators

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

$$F_{\mathbf{w}}(a_1, \dots, a_n) = \sum_{i=1}^n w_i b_i, \quad (6)$$

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

$$\text{orness}(\mathbf{w}) = \frac{1}{n-1} \sum_{i=1}^n (n-i)w_i, \quad (7)$$

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

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

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

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

$$F_{\mathbf{w}^*}(a_1, \dots, a_n) \geq \frac{1}{n} \sum_{i=1}^n a_i \quad (9a)$$

152 if  $\gamma \geq 0.5$  and

$$F_{\mathbf{w}^*}(a_1, \dots, a_n) \leq \frac{1}{n} \sum_{i=1}^n a_i \quad (9b)$$

153 if  $\gamma \leq 0.5$ .

### 154 2.3. Decision-making framework

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

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

171 To measure the desirability of consequences, we usually define a utility function  $U$  from  
172  $\mathcal{C}$  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  
173 general term  $u_{ij}$  of  $\mathbf{U}$  represents the utility of selecting act  $f_i$  (predicting the label as  $\omega_i$ )  
174 when the true class is  $\omega_j$ . In practice, matrix  $\mathbf{U}$  is often assumed to be the identity matrix,  
175 in which case the utility is equal to 1 for a correct decision, and 0 for a misclassification.  
176 However, the DM can also assign different utilities to different types of misclassification,  
177 making  $\mathbf{U}$  different from the identity matrix, and possibly asymmetric<sup>3</sup>.

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

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

### 198 3. Decision-making for evidential classification

199 Assuming the DM's information concerning the possible states of nature to be represented  
200 by a mass function  $m$  on  $\Omega$ , we now carry out a thorough analysis and comparison of different  
201 decision-making criteria to obtain label predictions, especially set-valued ones. A method  
202 for computing the utility of set-valued predictions is first introduced in Section 3.1. Precise  
203 classification with and without rejection is then recalled in Section 3.2, after which two  
204 main approaches to partial classification are studied: via complete preorder among partial  
205 assignments (Section 3.3) and via partial preorder among complete assignments (Section  
206 3.4). In Section 3.5, the time complexity issue is considered. Finally, the evaluation of  
207 set-based predictions is discussed in Section 3.6.

---

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



208 *3.1. Generating utilities of partial assignments via an OWA operator*

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

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

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

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

$$\bar{u}_{K,j} \leq \tilde{u}_{K,j} \leq \max_{\omega_i \in K} u_{ij} \quad (10)$$

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

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

$$\tilde{u}_{K,j} = F_{\mathbf{w}}(\{u_{ij} : \omega_i \in K\}) = \sum_{k=1}^{|K|} w_k u_{(k)j}^K, \quad (11)$$

236 where  $u_{(k)j}^K$  denotes the  $k$ -th largest element in the set  $\{u_{ij}, \omega_i \in K\}$ . In (11), each weight  
 237  $w_k$  can be interpreted as measuring the DM's preference to choose  $u_{(k)j}^K$  if forced to select  
 238 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  $\tilde{\mathbf{U}}$  extended by OWA operators with  $\gamma = 0.8$  and  $\gamma = 0.6$  (Example 1).

act	$\gamma = 0.8$			$\gamma = 0.6$		
	$\omega_1$	$\omega_2$	$\omega_3$	$\omega_1$	$\omega_2$	$\omega_3$
$f_{\{\omega_1\}}$	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

239 define the DM's *imprecision tolerance degree* as

$$\text{tol}(\mathbf{w}) = \sum_{k=1}^{|K|} \frac{|K| - k}{|K| - 1} w_k = \gamma. \quad (12)$$

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

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

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

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

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

### 256 3.2. Precise classification with and without rejection

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

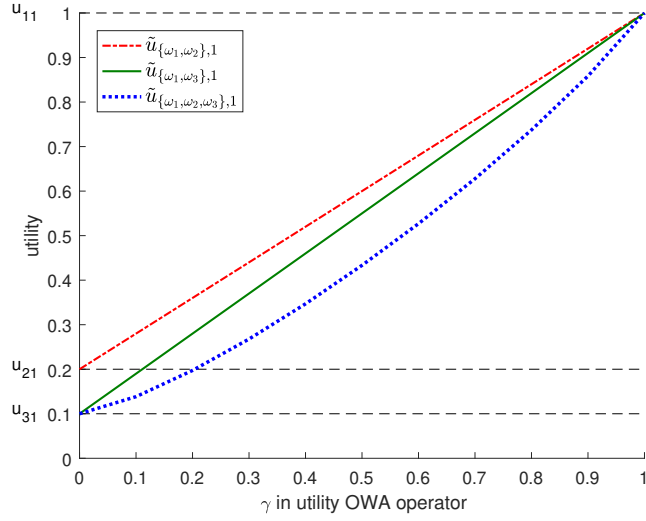


Figure 2: Aggregated utilities vs. imprecision tolerance degree  $\gamma$ .

259 decision criteria:

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

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

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

265 **Generalized maximax criterion.** Taking an optimistic point of view, a complete pre-  
 266 order can be defined from the upper expected utility

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

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

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

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

272 In [9], Denœux also proposed decision rules with rejection. Instances likely to be mis-  
 273 classified are rejected. Rejection can be identified with an additional act  $f_\Omega$  that consists in  
 274 assigning the instance to the whole set of classes. In [9], it was proposed to set  $u_{\Omega,j}$  to some  
 275 fixed value  $\lambda_0$  for all  $j$ .

276 In the next two sections, we extend the notion of rejection by allowing partial assignment  
 277 not only to  $\Omega$ , but also to any subset of  $\Omega$ .

### 278 3.3. Partial classification via complete preorders among partial assignments

279 As explained in Section 2.3, when considering partial assignments, the set of available  
 280 acts is  $\tilde{\mathcal{F}} = \{f_K, K \subseteq \Omega, K \neq \emptyset\}$ . A mass function  $m$  on  $\Omega$  and an extended utility  
 281 matrix  $\tilde{\mathbf{U}}_{(2^n-1) \times n}$  constructed, e.g., using the approach described in Section 3.1, are used for  
 282 decision-making. In addition to the previous three decision criteria, the following additional  
 283 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 com-  
 bination [41]. The generalized maximin and maximax criteria are special cases corre-  
 sponding, 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  $f_K \succ_\alpha f_G \iff \mathbb{E}_{m,\alpha}(f_K) \geq \mathbb{E}_{m,\alpha}(f_G)$   
 with

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

284 **Generalized OWA criterion.** Another generalization of the maximin and maximax cri-  
 285 teria consists in aggregating the utilities within each focal set  $K \subseteq \Omega$  using OWA  
 286 operators [47]. We have

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

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

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

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

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

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

acts	$\omega_1$	$\omega_2$	$\mathbb{E}_m(f_K)$	$\overline{\mathbb{E}}_m(f_K)$	$\mathbb{E}_p(f_K)$	$\mathbb{E}_{m,\alpha}(f_K)$	$\mathbb{E}_{m,\beta}^{owa}(f_K)$	$\overline{R}_K$
$f_{\{\omega_1\}}$	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

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

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

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

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

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

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

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

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

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

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

323 Since  $f_K \succ_* f_G \iff \mathbb{E}_m(f_K) \geq \mathbb{E}_m(f_G)$  and  $f_K \succ_r f_G \iff \bar{R}(f_K) \leq \bar{R}(f_G)$ , the complete  
324 preference relations induced by the lower expectation and by the expected maximal regret  
325 are identical. Therefore, these two decision criteria always reach the same decision.  $\square$

### 326 3.4. Partial classification via partial preorders among precise assignments

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

333 The decision-making criteria involved in this approach are more rooted in the imprecise  
334 probability view of belief functions [43, 13]. With insufficient information about states of na-  
335 ture, each precise assignment  $f_i$  corresponds to an expected utility interval  $[\underline{\mathbb{E}}_m(f_i), \overline{\mathbb{E}}_m(f_i)]$ ,  
336 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  
337 be viewed as the range of expectations  $\mathbb{E}_P(f)$  with respect to all probability measures  $P$   
338 compatible with mass function  $m$ . By comparing the acts in  $\mathcal{F}$  based on the lower and  
339 upper expected utilities, partial preorders of precise assignments are obtained using the fol-  
340 lowing decision criteria. (The reader is referred to [13] for more a detailed presentation and  
341 discussion of these criteria).

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

348 **Weak dominance criterion.** According to this less conservative criterion [13],  $f_i \succ_{WD} f_j$   
 349 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  $f_i \succ_{SD} f_j$  implies  
 350  $f_i \succ_{WD} f_j$ . The set of non-dominated elements of  $\succ_{WD}$  is included in that of  $\succ_{SD}$ ,  
 351 *i.e.*,  $\mathcal{F}_{WD}^* \subseteq \mathcal{F}_{SD}^*$ .

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

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

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

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

where  $\alpha_u$  and  $\beta_u$  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

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

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

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

$$\min_{\mathbf{a}, \mathbf{p}, \boldsymbol{\lambda}} \sum_{\ell \neq i} \lambda_\ell \quad (20a)$$

369 subject to

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

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

371

$$p_k = \sum_{j=1}^q a_{kj}, \quad k = 1, \dots, n \quad (20d)$$

372

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

373

$$\lambda_\ell \geq 0, \quad \ell \neq i, \quad (20f)$$

374

375

376

377

378

379

380

where  $F_1, \dots, F_q$  are the focal elements of  $m$ , vector  $\mathbf{a}$  contains all the allocations<sup>4</sup>  $a_{kj} = a(\omega_k, F_j)$  such that  $\omega_k \in F_j$ ,  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{i-1}, \lambda_{i+1}, \dots, \lambda_n)$  are  $n - 1$  slack variable vectors, and the vector  $\mathbf{p} = (p_1, \dots, p_n)$  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.

381

382

383

384

385

386

387

388

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.

389

390

391

392

393

394

395

396

397

398

399

400

**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  $\beta_u = 0$ .

401

402

403

404

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.

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



405 *3.5. Discussion on time complexity*

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

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

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

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

436 *3.6. Evaluation of set-valued predictions*

437 In classification applications, a test set  $\mathcal{T}$  is typically used to assess the performance of  
 438 a trained classifier. To choose a proper criterion, a standard is needed to evaluate different  
 439 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  $\omega$  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:

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

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

443 In this paper, we consider the case of utilities in  $[0, 1]$ . We aim to evaluate precise or  
 444 set-valued predictions by a single number (rather than a vector of parameters as proposed in  
 445 [5]), with the requirement that the better the prediction, the larger the performance measure.  
 446 In Section 3.1, we have seen how to extend a utility matrix so as to compute the utility of  
 447 partial assignments. Accordingly, we propose to evaluate the classification performance by  
 448 the *averaged utility* of the decisions made for the instances in test set  $\mathcal{T}$ :

$$\text{AvU}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \tilde{u}_{K(i), \omega(i)}. \quad (21)$$

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

#### 451 4. Experiments and discussions

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

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

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

#### 471 4.1. Iris data set

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

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

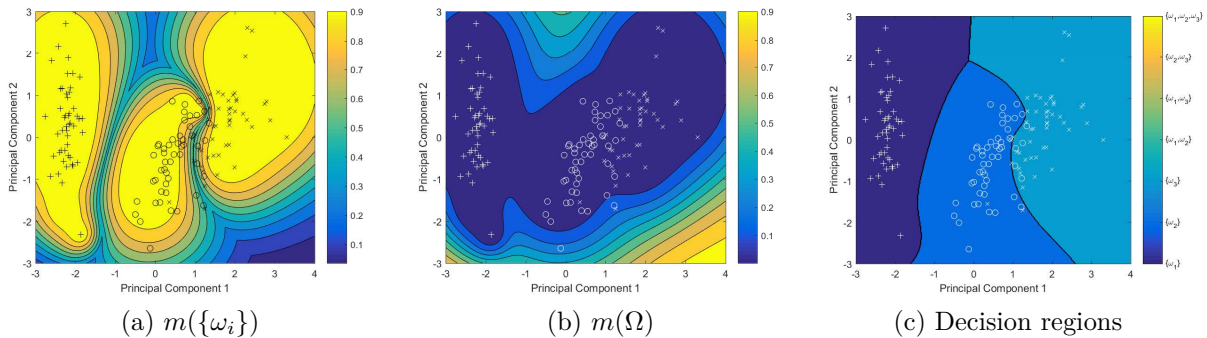


Figure 3: Iris data set: contours of mass functions provided by the evidential neural networks (a,b) and decision regions for the decision rules with precise assignment (c).

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

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

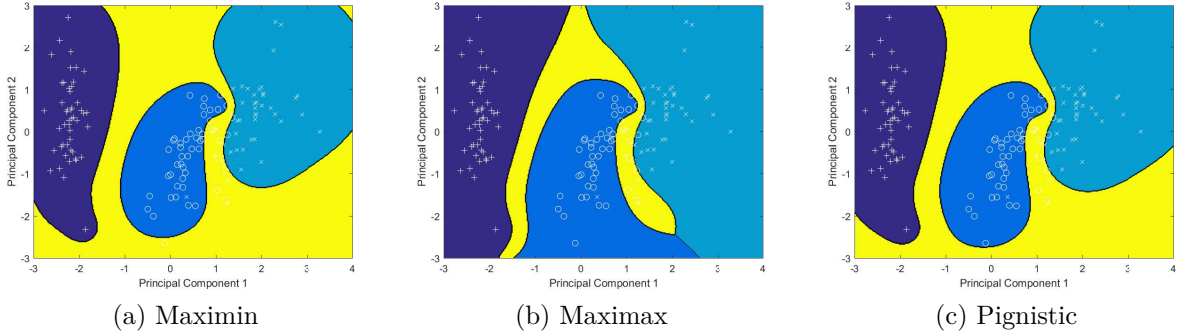


Figure 4: Iris data set: decision regions with rejection (obtained for  $\lambda_0 = 0.8$ ).

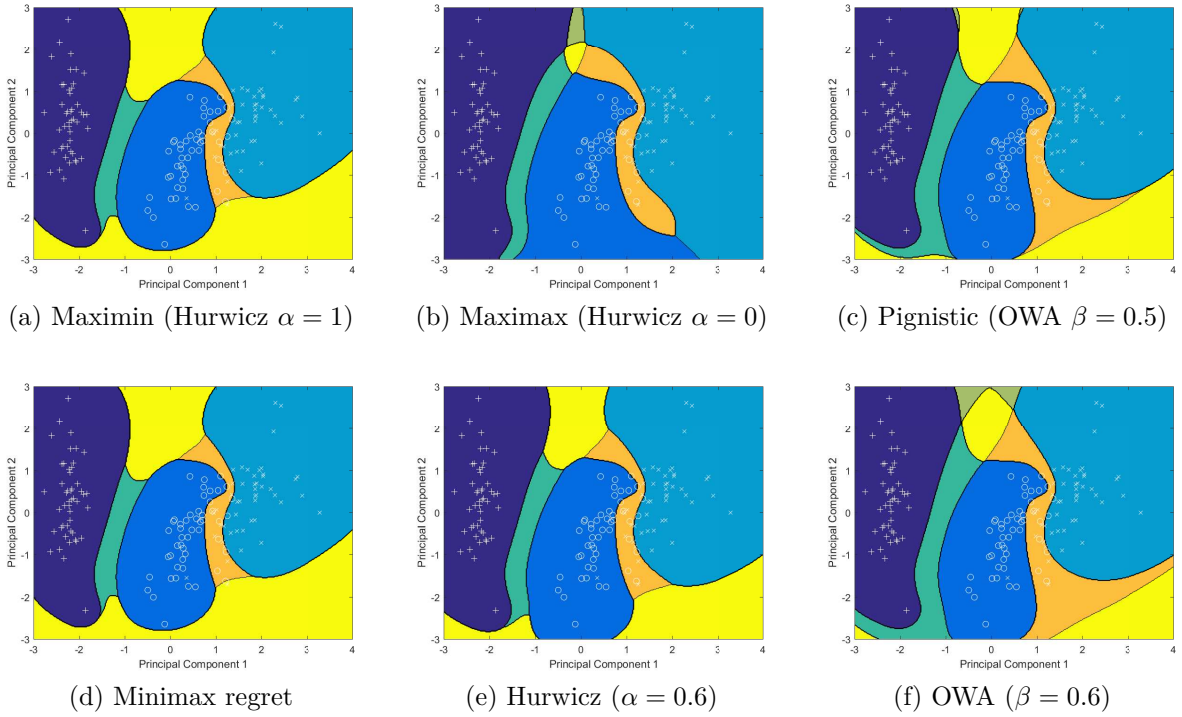


Figure 5: Iris data set: partial classification (complete preorder case) with the six decision criteria.

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

503 informative predictions are made by our approach.

504 Figure 6 shows the decision boundaries for the Iris data set obtained with partial pre-  
 505 orders among precise assignments (Section 3.4). For the interval-valued utility criterion, we  
 506 set  $\alpha_u = 0.7$  and  $\beta_u = 0.3$ . For this data set, there is no difference among some criteria. The  
 507 weak dominance and interval-valued utility criteria always give precise predictions, while the  
 508 other three criteria can differentiate conflict from lack of information. Compared to previ-  
 509 ous results based on complete preorders of partial assignments (Figure 5), fewer set-valued  
 510 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.

#### 511 4.2. Classification performances with varying $\gamma$

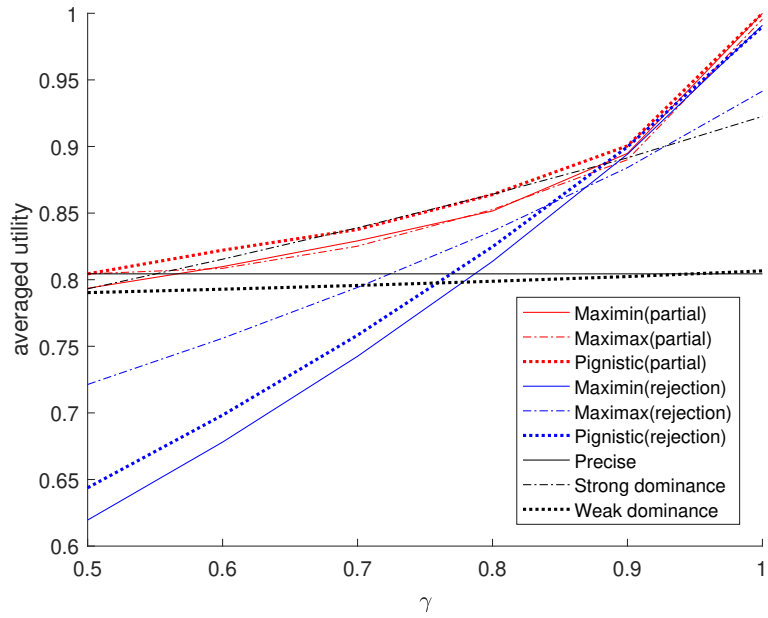
512 We then checked the averaged utilities obtained from utility matrices extended with  
 513 varying imprecision tolerance degree  $\gamma$ . Experiments were carried out using 22 data sets  
 514 from the UCI repository [25] with characteristics summarized in Table 3. The original  
 515 utility matrix  $\mathbf{U}$  was assumed to be the identity matrix of size  $n$  (the number of classes). The  
 516 evidential neural network was set with six prototypes per class, except for Lung cancer and  
 517 Annealing data sets with three prototypes per class. Attributes or instances with more than  
 518 30% missing values were removed. To evaluate the performances, five-fold cross-validation  
 519 was performed, and all experiments were repeated five times to compute an average result.

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

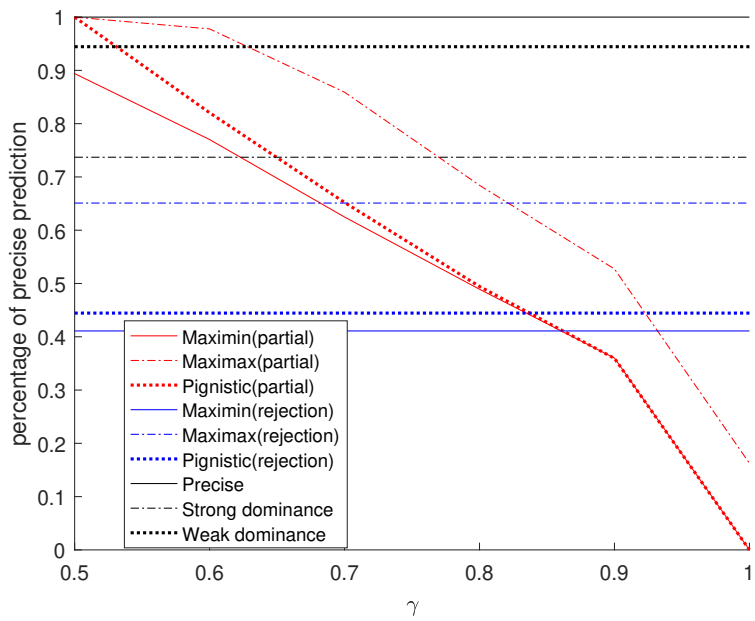
529 The averaged utilities and corresponding percentages of precise prediction for the Wine  
 530 and Breast tissue data sets are shown, respectively, in Figures 7 and 8. In general, partial  
 531 classification methods based on complete preorder (red curves) work better than those based

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

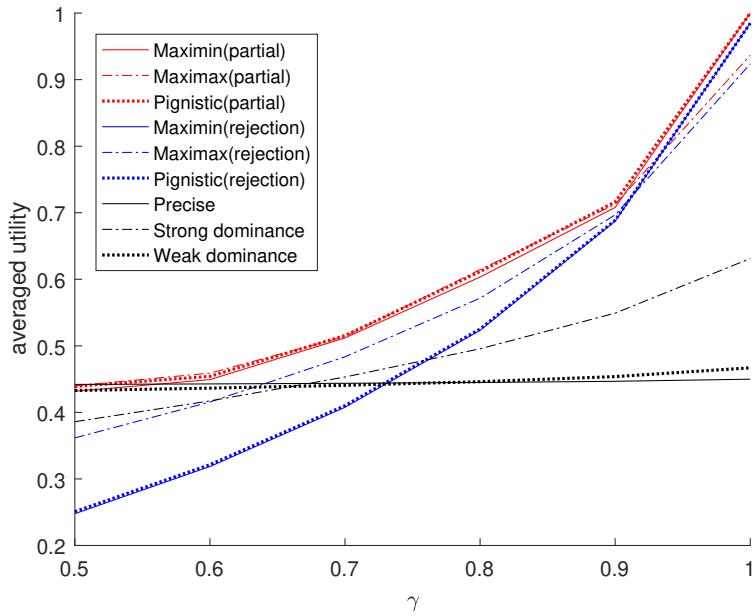


(a) Averaged utility

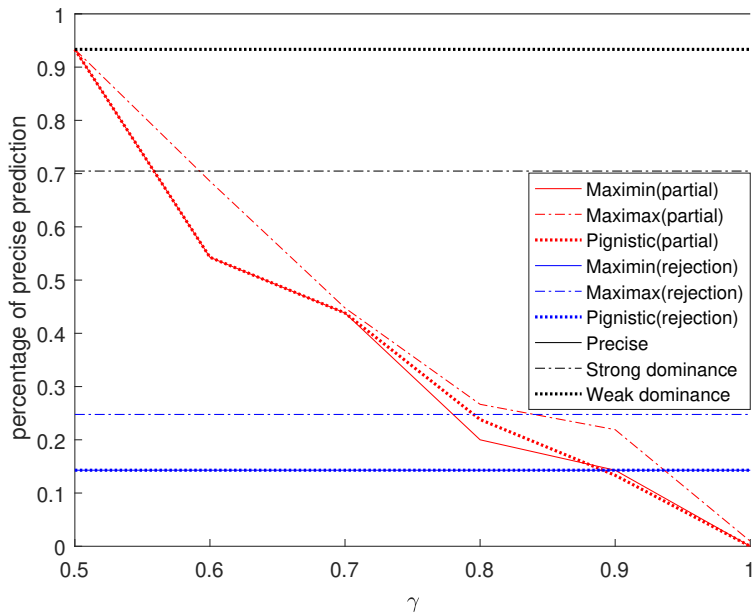


(b) Percentage of precise prediction

Figure 7: Experimental results with varying  $\gamma$  (Wine Data set).



(a) Averaged utility



(b) Percentage of precise prediction

Figure 8: Experimental results with varying  $\gamma$  (Breast tissue data set).



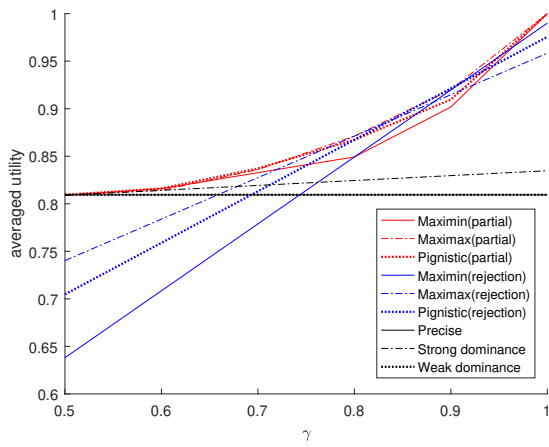
532 on partial preorder. They also perform better than rejection (blue curves). Compared to  
533 precise classifications, rejection methods have lower averaged utilities for small  $\gamma$  and higher  
534 ones for large  $\gamma$ . This general tendency can be explained as follows.

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

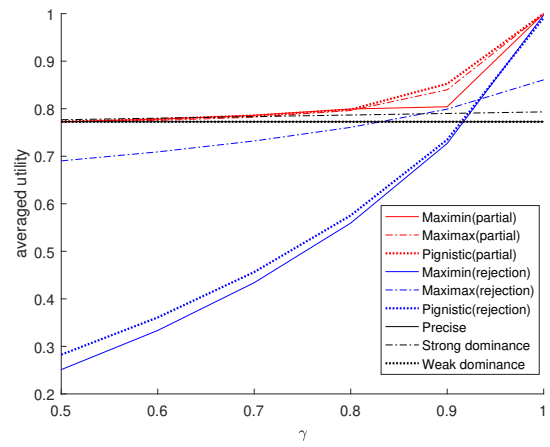
545 For partial classification based on complete preorders among partial assignments, given  
546 a fixed  $\gamma$ , the averaged utilities of different criteria vary in a small range, as shown by the  
547 red lines in Figures 7a and 8a. Among the Maximin, Maximax and Pignistic criteria, we can  
548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

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

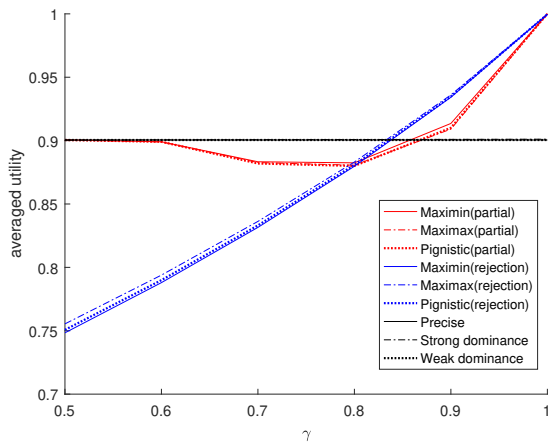
570 There is something special for some data sets such as Balance scale, Dermatology and Iris:  
571 Considering partial classification based on complete preorders (red curves in Figures 9c, 9f  
572 and 11c), the averaged utility curves have a U-shape (they first decrease when  $\gamma$  increases,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



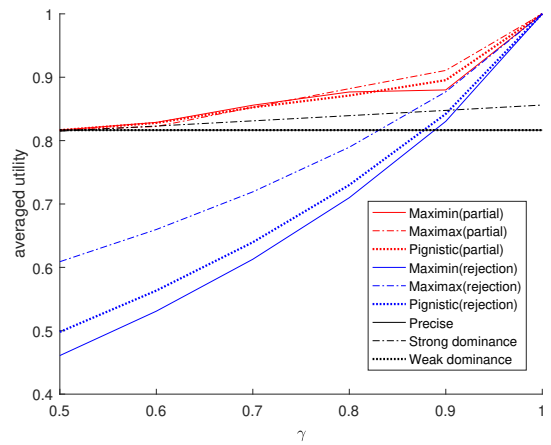
(a) Adult data set



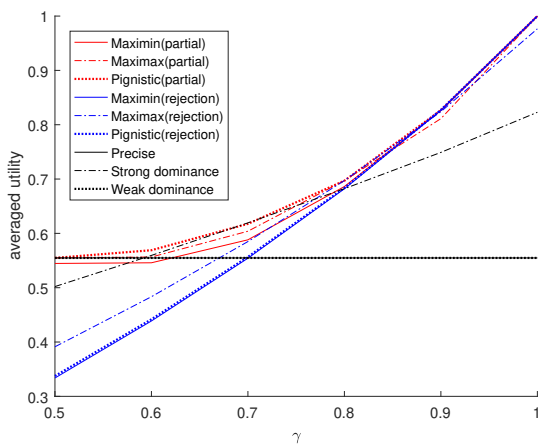
(b) Annealing data set



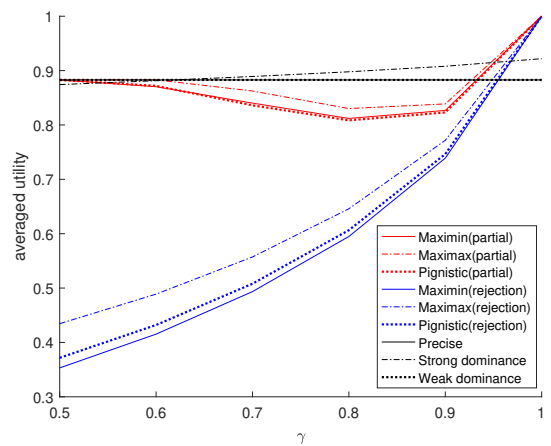
(c) Balance scale data set



(d) Car evaluation data set

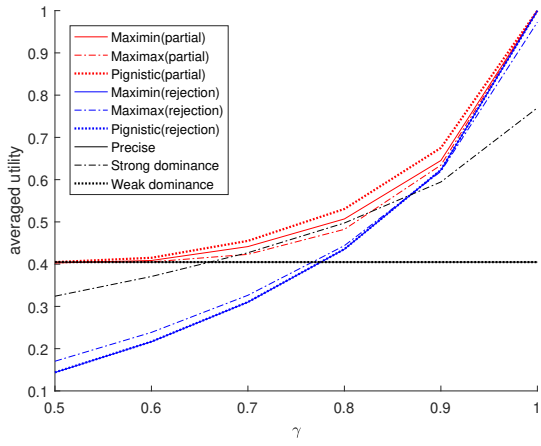


(e) Contraceptive method choice data set

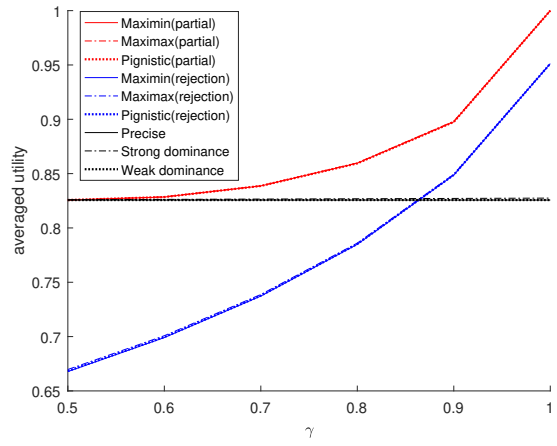


(f) Dermatology data set

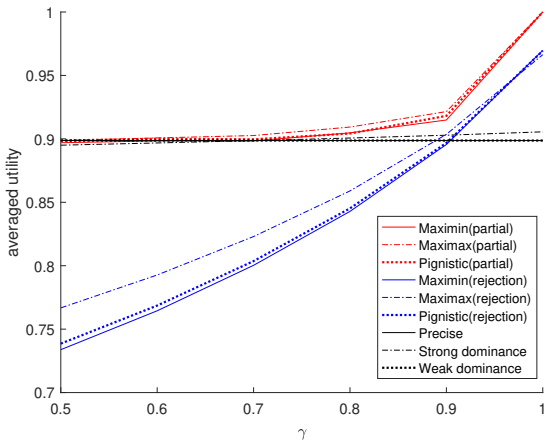
Figure 9: Averaged utilities with varying  $\gamma$  (part 1).



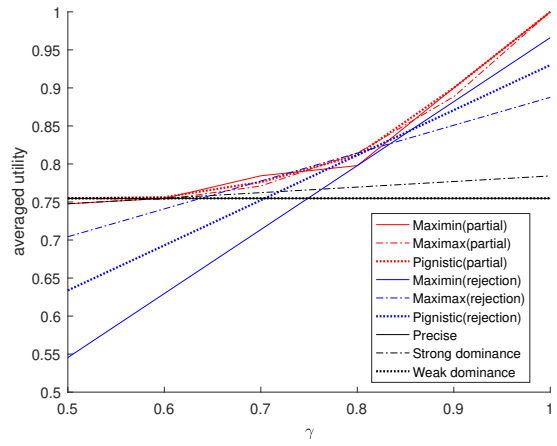
(a) Drug consumption data set



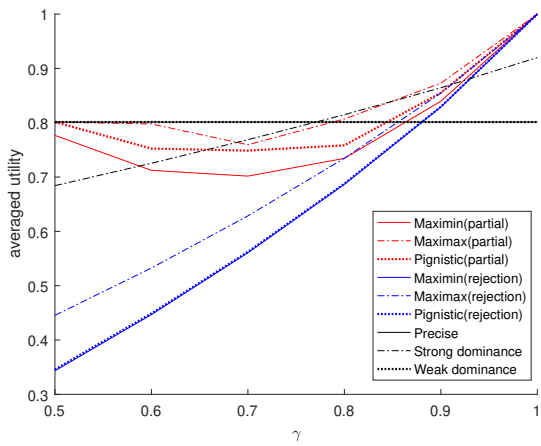
(b) Ecoli data set



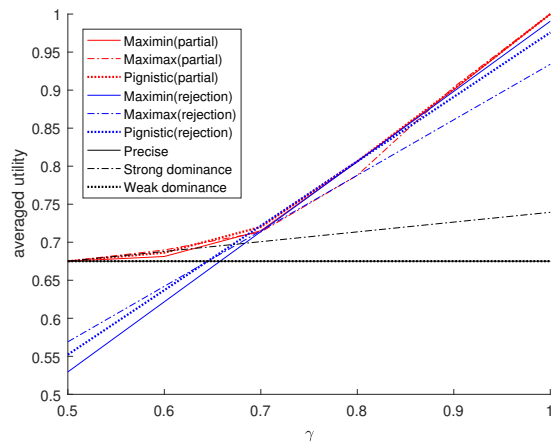
(c) Forest mapping data set



(d) Harberman's survival data set

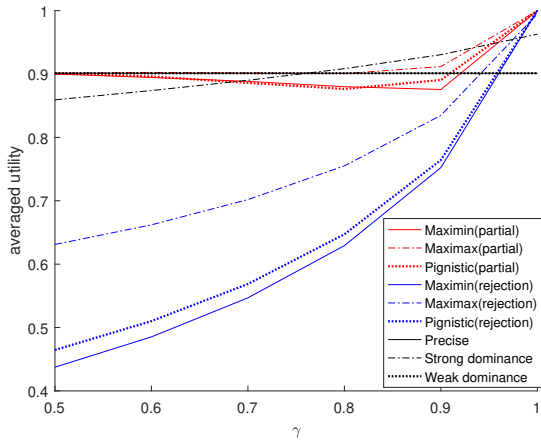


(e) Hayes-roth data set

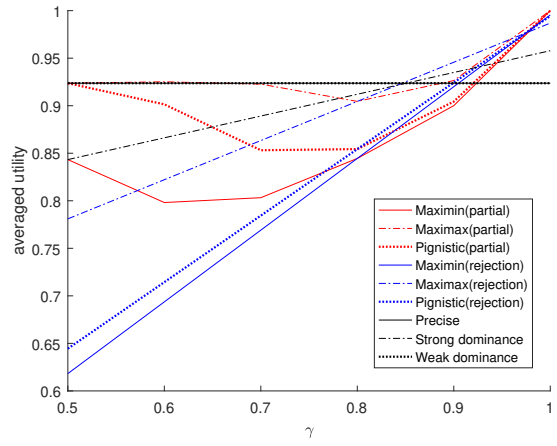


(f) Hepatitis data set

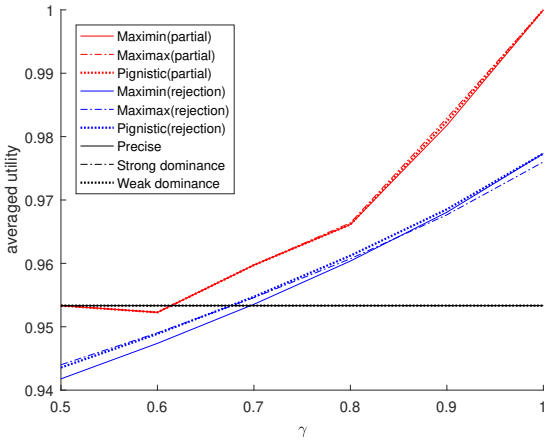
Figure 10: Averaged utilities with varying  $\gamma$  (part 2).



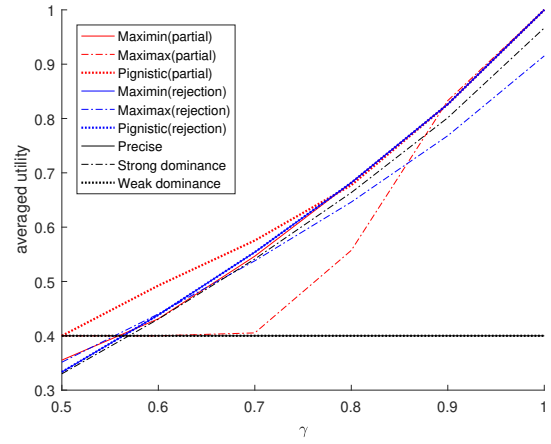
(a) Image segmentation data set



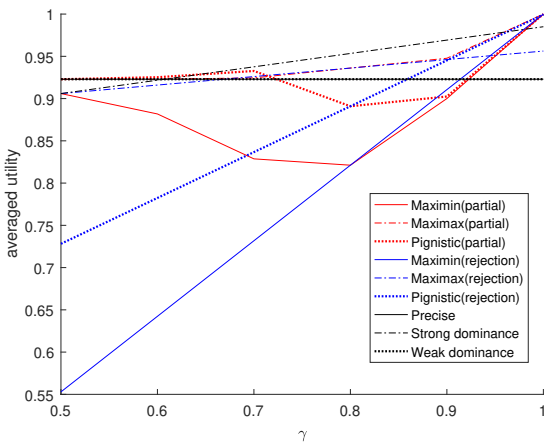
(b) Ionosphere data set



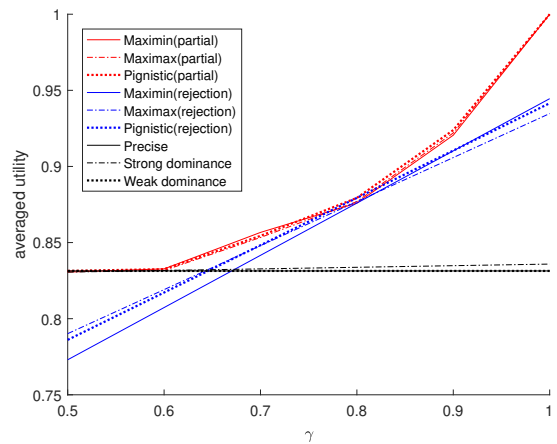
(c) Iris data set



(d) Lung cancer data set



(e) Mushroom data set



(f) SPECT heart data set

Figure 11: Averaged utilities with varying  $\gamma$  (part 3).

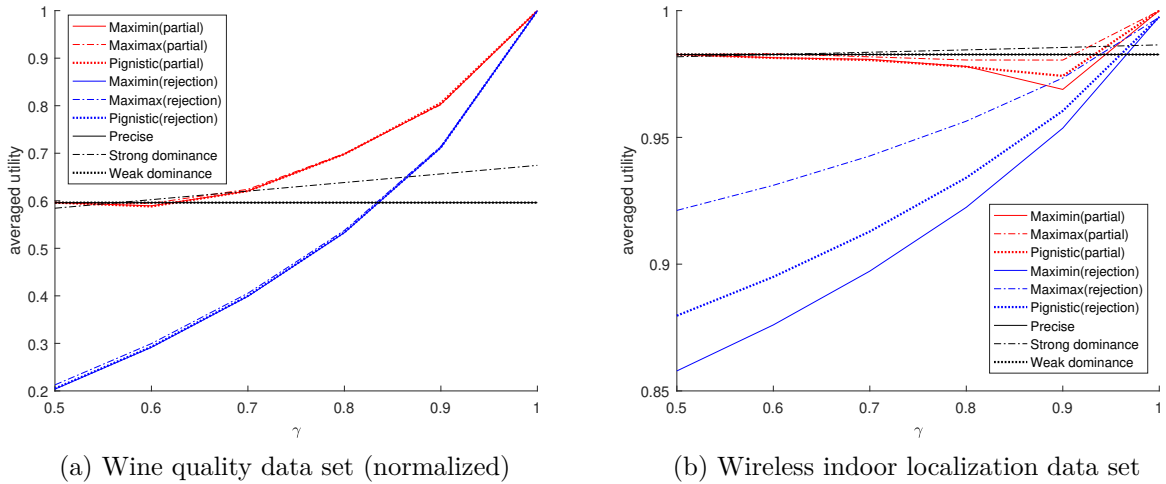


Figure 12: Averaged utilities with varying  $\gamma$  (part 4).

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

#### 585 4.3. Performances with Noisy Test Sets

586 In many situations, a classifier is trained with “good” data (acquired and preprocessed in  
587 controlled conditions) and then used in a real environment where, for instance, sensors may  
588 not be well calibrated. In such a case, the test data do not have the same distribution as the  
589 learning data. Cautious decision rules making set-valued predictions can be expected to be  
590 particular beneficial in such an environment, and discrepancies between the performances  
591 of different decisions rules may be more apparent than they are in the case of “clean” data  
592 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:

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

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

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

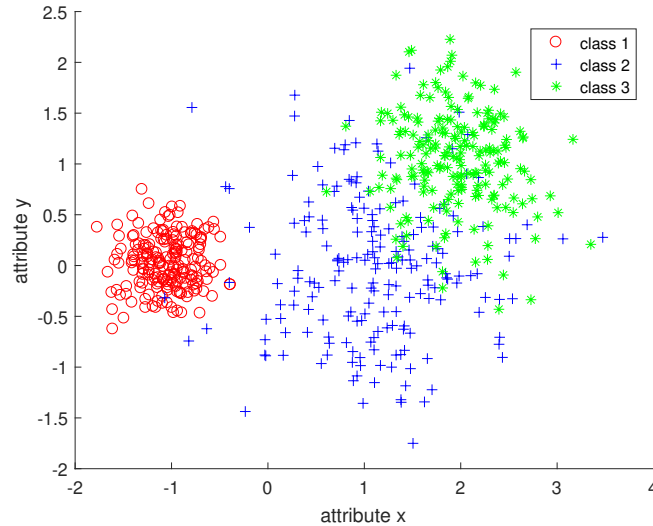


Figure 13: A Gaussian data set of 600 instances.

---

**Algorithm 1:** Algorithm to generate a noisy test set.

---

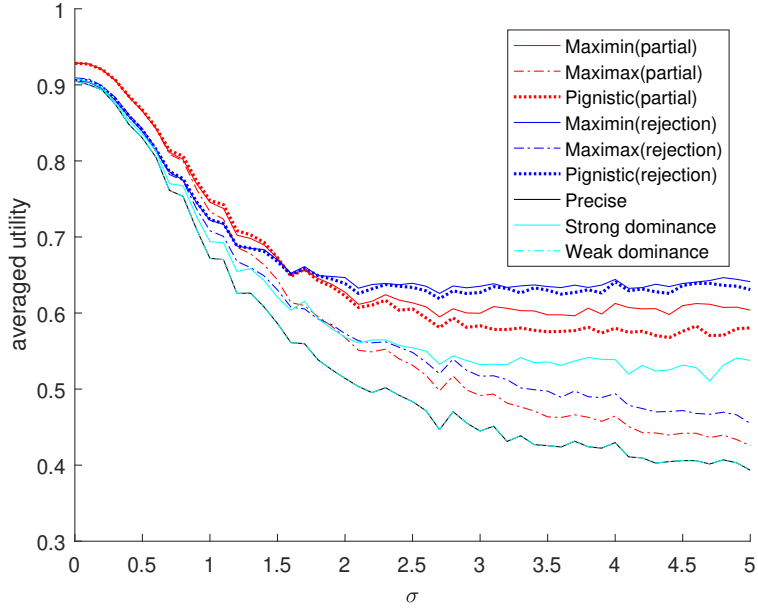
**Input:** test set  $T = \{(\mathbf{x}, \mathbf{y}), C\}$ , noise standard deviation  $\sigma$

**Output:** noisy test set  $\tilde{T} = \{(\tilde{\mathbf{x}}, \mathbf{y}), C\}$

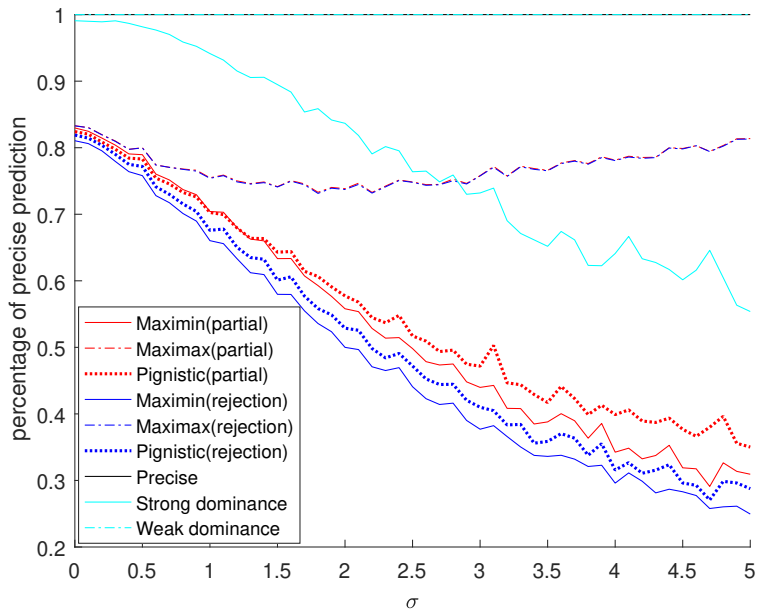
- 1 **for**  $1 \leq i \leq |T|$  **do**
  - 2     Draw  $\epsilon(i)$  from  $\mathcal{N}(0, \sigma^2)$ ;
  - 3      $\tilde{x}(i) = x(i) + \epsilon(i)$ ;
- 

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

604 For  $\sigma = 0$ , the test and training sets have the same distribution; partial classification  
 605 via complete preorders among partial assignments (Section 3.3) outperforms other partial  
 606 classification approaches (Section 3.4) as well as precise classifications with and without  
 607 rejection slightly. When  $\sigma$  increases, the averaged utilities for all criteria drop quickly  
 608 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.

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

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

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

#### 628 4.4. Parameter selection of decision criteria

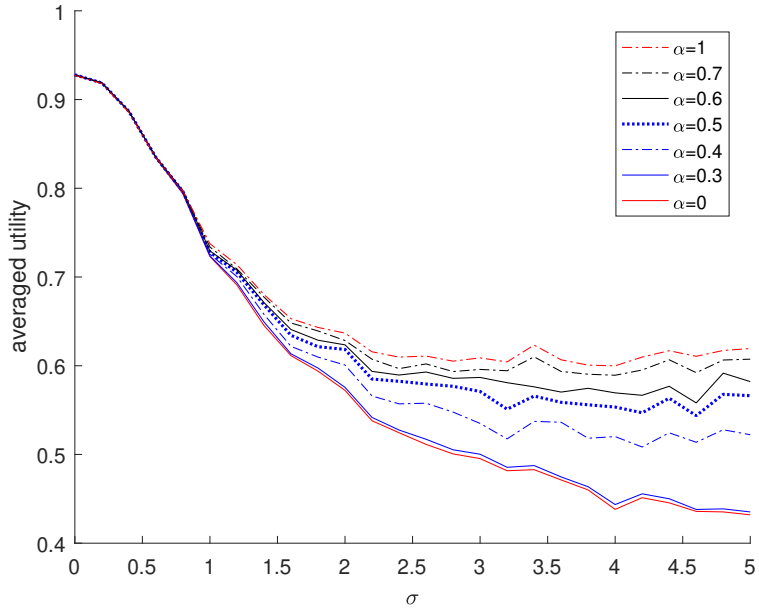
629 In the fourth experiment, we study the problem of parameter selection in several decision  
 630 criteria. Consider again the simulated Gaussian data set, we kept the same data setting as  
 631 in Section 4.3 and we set  $\gamma = 0.8$ . Figures 15a and 15b show, respectively, the decisions  
 632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

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

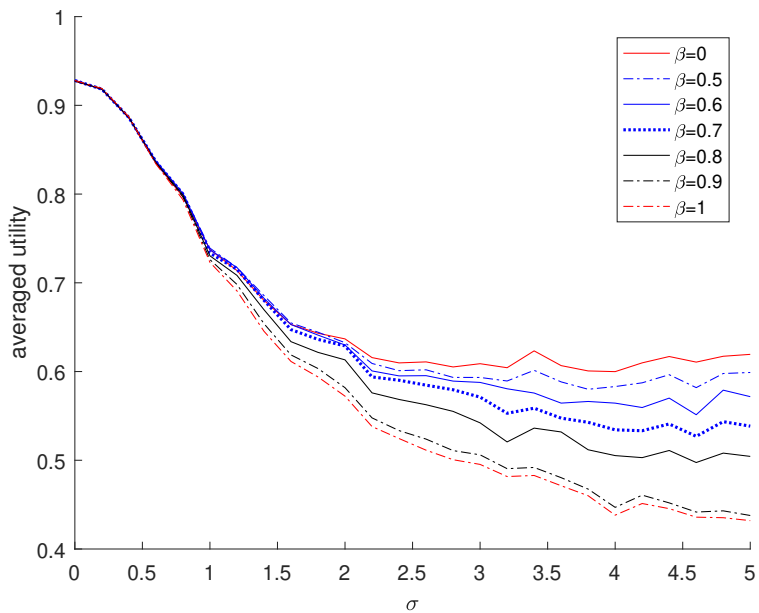
640 For the OWA decision criterion, the averaged utility decreases monotonically as  $\beta$  grows.  
 641 When  $\beta = 0.5$  (pignistic criterion), the performance is quite acceptable. The further decrease  
 642 of  $\beta$  leads to little improvement of performance. On the other side, the averaged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

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





(a) Hurwicz criterion



(b) OWA criterion

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

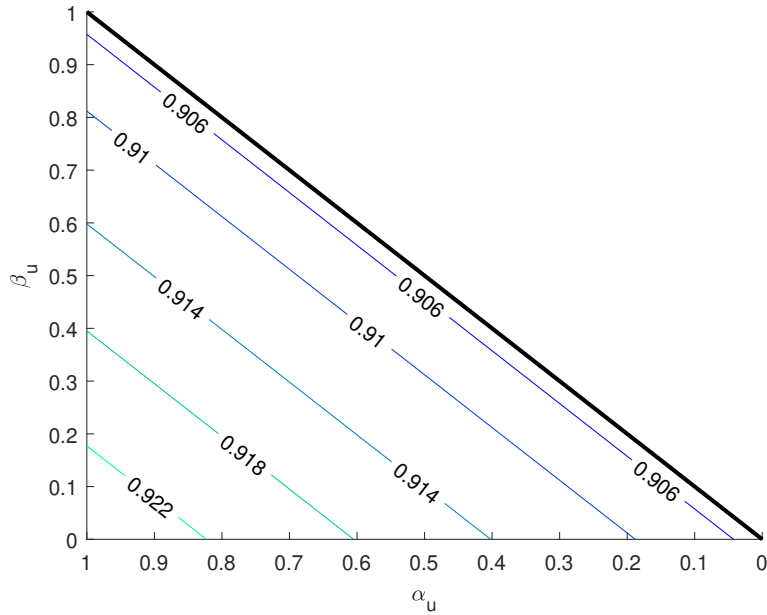


Figure 16: Averaged utilities with changing  $\alpha_u$  and  $\beta_u$ .

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

## 657 5. Conclusion

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

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

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

## 680 Acknowledgement

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

## 685 Appendix A. Evidential neural network classifier

686 The *evidential neural network classifier* [10] is an adaptive classifier based on DS theory<sup>5</sup>.  
 687 The classification procedure can be implemented in a specific neural network architecture  
 688 with one input layer, two hidden layers and one output layer. Assessing the similarity  
 689 of a pattern with a limited number of prototypes, items of evidence regarding the class  
 690 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  $\mathbf{p}^1, \dots, \mathbf{p}^k$  are considered instead of the nearest neighbors of  
 a pattern  $\mathbf{x}$  to reduce the computational complexity. Each prototype  $\mathbf{p}^i$  provides a mass  
 function  $m^i$  based on the Euclidean distance  $d^i = \|\mathbf{x} - \mathbf{p}^i\|$  between  $\mathbf{x}$  and  $\mathbf{p}^i$ ,  $i = 1, \dots, k$ .  
 The unit mass is distributed among the singleton  $\{\omega_q\}$  and  $\Omega$ :

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

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

691 where  $\gamma^i$  is a scale parameter for prototype  $\mathbf{p}^i$ ,  $u_q^i$  is the membership degree of prototype  $\mathbf{p}^i$   
 692 to class  $\omega_q$  (with  $\sum_{q=1}^n u_q^i = 1$ ), and  $\alpha^i$  is a parameter indicating the relative importance of  
 693 prototype  $\mathbf{p}^i$  in classifying new patterns. Combining the  $k$  mass functions  $m^i, i = 1, \dots, k$   
 694 by Dempster’s rule (4), the resulted mass function describes the uncertainty pertaining to  
 695 the class of a pattern.

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

---

<sup>5</sup>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](https://www.hds.utc.fr/~tdenoeux/dokuwiki/en/software/belief_nn).

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

- 702 [1] K. Ali, S. Manganaris, and R. Srikant. Partial classification using association rules. In *KDD'97: Pro-*  
703 *ceedings of the Third International Conference on Knowledge Discovery and Data Mining*, volume 97,  
704 pages 115–118, 1997.
- 705 [2] A. Appriou. Uncertain data aggregation in classification and tracking processes. In B. Bouchon-Meunier,  
706 editor, *Aggregation and Fusion of imperfect information*, pages 231–260. Physica-Verlag, Heidelberg,  
707 1998.
- 708 [3] T. Augustin, F. P. Coolen, G. De Cooman, and M. C. Troffaes. *Introduction to imprecise probabilities*.  
709 John Wiley & Sons, 2014.
- 710 [4] C.-K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on*  
711 *Electronic Computers*, EC-6(4):247–254, 1957.
- 712 [5] G. Corani and M. Zaffalon. Lazy naive credal classifier. In *Proceedings of the 1st ACM SIGKDD*  
713 *Workshop on Knowledge Discovery from Uncertain Data*, pages 30–37. ACM, 2009.
- 714 [6] J. J. Del Coz, J. Díez, and A. Bahamonde. Learning nondeterministic classifiers. *Journal of Machine*  
715 *Learning Research*, 10(79):2273–2293, 2009.
- 716 [7] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The annals of*  
717 *mathematical statistics*, pages 325–339, 1967.
- 718 [8] T. Denœux. A  $k$ -nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans.*  
719 *on Systems, Man and Cybernetics*, 25(05):804–813, 1995.
- 720 [9] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern recognition*,  
721 30(7):1095–1107, 1997.
- 722 [10] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on*  
723 *Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.
- 724 [11] T. Denœux. 40 years of Dempster-Shafer theory. *International Journal of Approximate Reasoning*,  
725 79:1–6, 2016.
- 726 [12] T. Denœux. *evclass: Evidential Distance-Based Classification*, 2017. R package version 1.1.1.
- 727 [13] T. Denœux. Decision-making with belief functions: A review. *International Journal of Approximate*  
728 *Reasoning*, 109:87–110, 2019.
- 729 [14] T. Denœux. Logistic regression, neural networks and Dempster-Shafer theory: A new perspective.  
730 *Knowledge-Based Systems*, 176:54–67, 2019.
- 731 [15] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Beyond  
732 probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial*  
733 *Intelligence Research*, volume 1, chapter 4, pages 119–150. Springer Verlag, 2020.
- 734 [16] T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta. A new evidential  $k$ -nearest neighbor rule based  
735 on contextual discounting with partially supervised learning. *International Journal of Approximate*  
736 *Reasoning*, 113:287–302, 2019.
- 737 [17] T. Denœux and P. P. Shenoy. An interval-valued utility theory for decision making with Dempster-  
738 Shafer belief functions. *International Journal of Approximate Reasoning*, 124:194–216, 2020.
- 739 [18] D. Filev and R. R. Yager. Analytic properties of maximum entropy OWA operators. *Information*  
740 *Sciences*, 85(1):11–27, 1995.
- 741 [19] E. Frank and M. Hall. A simple approach to ordinal classification. In *European Conference on Machine*  
742 *Learning*, pages 145–156. Springer, 2001.
- 743 [20] G. Fumera, F. Roli, and G. Giacinto. Reject option with multiple thresholds. *Pattern recognition*,  
744 33(12):2099–2101, 2000.
- 745 [21] T. M. Ha. The optimum class-selective rejection rule. *IEEE Transactions on Pattern Analysis and*  
746 *Machine Intelligence*, 19(6):608–615, 1997.

- 747 [22] R. Herbei and M. H. Wegkamp. Classification with reject option. *Canadian Journal of Statistics*,  
748 34(4):709–721, 2006.
- 749 [23] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep  
750 feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning*  
751 *Systems*, 29(8):3573–3587, 2017.
- 752 [24] I. Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT  
753 press, 1983.
- 754 [25] M. Lichman. UCI machine learning repository, 2013.
- 755 [26] X. Liu and L. Chen. On the properties of parametric geometric OWA operator. *International Journal*  
756 *of Approximate Reasoning*, 35(2):163–178, 2004.
- 757 [27] Z.-G. Liu, L. Huang, K. Zhou, and T. Denœux. Combination of transferable classification with multi-  
758 source domain adaptation based on evidential reasoning. *IEEE Transactions on Neural Networks and*  
759 *Learning Systems*, pages 1–15, 2020.
- 760 [28] Z.-G. Liu, Q. Pan, J. Dezert, and A. Martin. Combination of classifiers with optimal weight based on  
761 evidential reasoning. *IEEE Transactions on Fuzzy Systems*, 26(3):1217–1230, 2018.
- 762 [29] Z.-G. Liu, Q. Pan, G. Mercier, and J. Dezert. A new incomplete pattern classification method based  
763 on evidential reasoning. *IEEE Transactions on Cybernetics*, 45(4):635–646, 2015.
- 764 [30] L. Ma and T. Denœux. Making set-valued predictions in evidential classification: A comparison of  
765 different approaches. In *International Symposium on Imprecise Probabilities: Theories and Applications*  
766 *(ISIPTA 2019)*, pages 276–285, Ghent, Belgium, July 2019. Published in Proceedings of Machine  
767 Learning Research 103:276–285, 2019.
- 768 [31] T. Mortier, M. Wydmuch, K. Dembczyński, E. Hüllermeier, and W. Waegeman. Efficient set-valued  
769 prediction in multi-class classification. *arXiv preprint arXiv:1906.08129*, 2019.
- 770 [32] M. O’Hagan. Aggregating template or rule antecedents in real-time expert systems with fuzzy set logic.  
771 In *Twenty-Second Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 681–689,  
772 1988.
- 773 [33] E. Ramasso and R. Gouriveau. Prognostics in switching systems: Evidential markovian classification  
774 of real-time neuro-fuzzy predictions. In *2010 Prognostics and System Health Management Conference*,  
775 pages 1–10. IEEE, 2010.
- 776 [34] A. P. Reynolds and B. De la Iglesia. A multi-objective grasp for partial classification. *Soft Computing*,  
777 13(3):227–243, 2009.
- 778 [35] M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels.  
779 *Journal of the American Statistical Association*, 114(525):223–234, 2019.
- 780 [36] L. J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–  
781 67, 1951.
- 782 [37] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, 1976.
- 783 [38] G. Shafer. Constructive decision theory. *International Journal of Approximate Reasoning*, 79:45–62,  
784 2016.
- 785 [39] P. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International*  
786 *Journal of Approximate Reasoning*, 38(2):133–147, 2005.
- 787 [40] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66(2):191–243, 1994.
- 788 [41] T. M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*,  
789 4(5–6):391–417, 1990.
- 790 [42] Z.-G. Su, T. Denœux, Y.-S. Hao, and M. Zhao. Evidential K-NN classification with enhanced perfor-  
791 mance via optimizing a class of parametric conjunctive t-rules. *Knowledge-Based Systems*, 142:7–16,  
792 2018.
- 793 [43] M. C. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal*  
794 *of Approximate Reasoning*, 45(1):17–29, 2007.
- 795 [44] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science &  
796 Business Media, 2005.
- 797 [45] P. Walley. *Statistical reasoning with imprecise probabilities*. Monographs on statistics and applied

- 798 probability. Chapman and Hall, 1991.
- 799 [46] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision-making.  
800 *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190, 1988.
- 801 [47] R. R. Yager. Decision making under Dempster-Shafer uncertainties. *International Journal of General*  
802 *Systems*, 20(3):233–245, 1992.
- 803 [48] R. R. Yager. Decision making using minimization of regret. *International Journal of Approximate*  
804 *Reasoning*, 36(2):109–128, 2004.
- 805 [49] G. Yang, S. Destercke, and M.-H. Masson. The costs of indeterminacy: How to determine them? *IEEE*  
806 *Transactions on Cybernetics*, 47(12):4316–4327, 2017.
- 807 [50] M. Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1):5–21, 2002.
- 808 [51] M. Zaffalon, G. Corani, and D. Mauá. Evaluating credal classifiers by utility-discounted predictive  
809 accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.