Making Set-valued Predictions in Evidential Classification: A Comparison of Different Approaches

Liyao Ma CSE_MALY@UJN.EDU.CN
School of Electrical Engineering, University of Jinan, Jinan, China
Université de Technologie de Compiègne, CNRS UMR 7253 Heudiasyc, Compiègne, France

Thierry Denœux THIERRY.DENOEUX@UTC.FR
Université de Technologie de Compiègne, CNRS UMR 7253 Heudiasyc, Compiègne, France

Abstract

In classification, it is often preferable to assign a pattern to a set of classes when the uncertainty is too high to make a precise decision. In this paper, we consider the problem of making set-valued predictions in classification tasks, when uncertainty is described by belief functions. Two approaches are contrasted. In the first one, an act is defined as the assignment to only one class, and we define a partial preorder among acts. The set of non-dominated acts is then given as the prediction. In the second approach, an act is defined as the assignment to a set of classes, and we construct a complete preorder among acts. The two approaches are discussed and compared experimentally. A critical issue both to make decisions and to evaluate decision rules is to define the utility of set-valued prediction. To this end, we propose to model the decision maker's attitude towards imprecision using an Ordered Weighted Average (OWA) operator, which allows us to extend the utility matrix. An experimental comparison of different decision rules is performed using UCI and artificial Gaussian data sets.

Keywords: Belief functions, Dempster-Shafer theory, Decision under uncertainty

1. Introduction

In classification problems, given a model learned from the training set, decisions are made on predicting the labels of new instances [9]. Unfortunately, it sometimes happens that our belief about the states of nature cannot be modeled by a probability measure as the information is insufficient to identify precise probabilities [4]. In such a situation, the commonly-used Maximum Expected Utility principle (MEU) fails to give an adequate decision of label assignment. In this paper, decision strategies in the Dempster-Shafer (DS) framework [2, 14, 8] are discussed.

As reviewed in paper [7], various classical Bayesian decision rules and imprecise probability decision rules have been extended to the DS framework. Yet in classification applications investigated so far, given uncertain information about the states of nature by belief functions, decisions have mainly been made by simple means such as using pignistic probability [13] put forward by Smets [15], or selecting the class with highest mass value [11]. Taking better advantage of uncertainty, Denœux [3] analysed three principled decision rules with rejection in the DS framework. A further step is made in this paper, in which different decision approaches allowing for set-valued predictions are discussed. A problem arising from such kind of predictions is how to evaluate their performances. Zaffalon et al. [21] analysed the $\{0,1\}$ reward case and proposed an evaluation metric that takes the decision maker's degree of risk aversion into account. Considering a more general case, Yang et al. [20] provided some properties that the utility of set-valued prediction should follow and proposed the p-discounted costs method. In this paper, inspired by the ordered weighted average (OWA) operator [17], we propose another approach to define the utilities for all set-valued predictions based on the utility of precise ones.

Let us denote by $\Omega = \{\omega_1, \cdots, \omega_n\}$ the set of classes (states of nature). For classification problems, in general an act is defined as the assignment of an instance to one and only one of the n classes. The set of acts is $\mathscr{F} = \{f_{\omega_1}, \cdots, f_{\omega_n}\}$, where f_{ω_i} (or f_i for short) denotes the assignment to class ω_i. To make decisions, we define a utility matrix $\mathbb{U}_{n\times n}$, whose general term $u_{ij} \in [0,1]$ denotes the utility of selecting class ω_i when the true class is ω_j. The more desirable is the prediction, the higher utility it achieves. Without loss of generality, we assume that $u_{ii} = 1$ for all i (correct predictions all have unit maximum utility) throughout this paper. When uncertainty is described by a probability distribution on Ω, we can compute an expected utility for each act. A complete preference relation among all available alternatives $f \in \mathscr{F}$ can then be computed and the optimal act provides a precise prediction of the instance. However, in situations of uncertainty, it may be preferable to assign a pattern to a set of classes. Different decision strategies are available for that purpose when uncertainty is described using either imprecise probabilities, or Dempster-Shafer belief functions. In this paper, we focus on the latter model, but some ideas can be transposed to other models. Describing the decision maker's (DM) information con-

cerning the states of nature by a belief function m on Ω, we analyse how to make set-valued classification predictions by different ways.

2. Two Families of Set-valued Decision Strategies

We have seen that a complete preorder among precise assignments can be used to make precise predictions. To compute set-valued predictions under uncertainty, we can basically start by modifying either the preference relations or the acts. Different decision criteria allowing us to derive either partial preorders among precise assignments, or complete preorders among partial assignments are discussed below.

2.1. Partial Preorders Among Precise Assignments

In this approach, we still define the acts as precise assignments (assigning the instance to one and only one of the n classes). Due to lack of information, each act f_i induces lower and upper expected utilities defined, respectively, as

$$\underline{\mathbb{E}}_m(f_i) = \sum_{B\subseteq\Omega} m(B) \min_{\omega_j\in B} u_{ij} \quad \text{and}$$
$$\overline{\mathbb{E}}_m(f_i) = \sum_{B\subseteq\Omega} m(B) \max_{\omega_j\in B} u_{ij}.$$

It is well known that the interval $[\underline{\mathbb{E}}_m(f_i), \overline{\mathbb{E}}_m(f_i)]$ is also the range of expectations $\mathbb{E}_P(f_i)$ with respect to all probability measures P compatible with belief function m (called the *credal set* of m). In the imprecise probability framework, three main decision criteria have been proposed [16]: interval dominance, maximality and weak dominance. These criteria are described in Table 1. By comparing the lower and upper expected utilities of acts, these criteria produce partial preorders, in which some acts are incomparable but more desirable than the others. Therefore, to choose the best act, we drop some sub-optimal alternatives and obtain an optimal set $\mathscr{F}^*$ such that $\forall f_i, f_j \in \mathscr{F}^*,\ f_i \sim f_j$ and $\forall f_i \in \mathscr{F}^*,\ \forall f_j \notin \mathscr{F}^*,\ f_i \succ f_j$. For instance, if we have the partial relation $f_{\omega_1} \succ f_{\omega_3}$ and $f_{\omega_2} \succ f_{\omega_3}$ ($|\Omega| = 3$), then the optimal set is $\mathscr{F}^* = \{f_{\omega_1}, f_{\omega_2}\}$ and the set of predicted classes is $\{\omega_1,\ \omega_2\}$.

Note that we obtain partial preorders because some alternatives cannot be ordered without additional information: from an imprecise probability theory point of view, when the credal set is too large or, from the belief function point of view, when we have some masses on non-singleton focal elements. Given additional information about states of nature, we can further narrow the expected utility intervals and make them comparable. The more information we have, the smaller is the optimal set $\mathscr{F}^*$. When the belief functions become Bayesian, a complete preorder revealing a precise prediction is obtained.

2.2. Complete Preorders Among Partial Assignments

Let us now consider the other approach to make set-valued predictions by extending the set of acts. Here, we generalize the acts as partially assigning the instance to a non-empty subset A of Ω. The set of acts becomes $\mathscr{F} = \{f_A, A \in 2^\Omega \setminus \{\emptyset\}\}$, where 2^Ω denotes the power set of Ω. Obviously, to make decisions, the original utility matrix $\mathbb{U}_{n\times n}$ needs to be extended to $\hat{\mathbb{U}}_{(2^n-1)\times n}$, with each element $\hat{u}_{A,j}$ representing the utility of assigning an instance to the set A of classes when the true class is ω_j. The details about utility matrix extension will be discussed in Section 3.

Since each act now corresponds to a set of classes, to make set-valued predictions, a complete preorder among partial assignments should be defined. Assuming mass function m on Ω and the extended utility matrix $\hat{\mathbb{U}}_{(2^n-1)\times n}$ to be given, Table 2 recalls extensions of classical decision criteria inducing complete preorders of partial assignments f_A. Here only the basic information needed to achieve preorders is described due to length limitation; more complete descriptions of all decision criteria used can be found in [7]. Some additional remarks about Table 2 are the following: i) For the pignistic criterion [15], $BetP(\{\omega_j\}) = \sum_{\omega_j\in A} \frac{m(A)}{|A|}$ is the pignistic probability where $|A|$ denotes the cardinality of subset $A \subseteq \Omega$; ii) In the generalized OWA criterion, F_β is the maximum entropy OWA operator with orness β [18] (more details about the OWA operator are given in Section 3); iii) For the generalized minimax regret criterion, $r_{A_i,j} = \max_k \hat{u}_{A_k,j} - \hat{u}_{A_i,j}$ is the regret that act f_{A_i} is selected when the true state ω_j occurs [19].

Also, we can remark that MEU works as a special case where uncertainty about Ω is quantified by probabilities $p_1, \cdots, p_n$. In addition, it can be proved that, for any act f_A, the sum of $\underline{\mathbb{E}}_m(f_A)$ and $\overline{R}(f_A)$ always equals 1 in our utility settings. The complete preorder achieved by descending order of $\underline{\mathbb{E}}_m(f_\cdot)$ is the same as that achieved by ascending order of $\overline{R}(f_\cdot)$. Therefore, the maximin and miminax regret criteria always result in the same decision for our problem.

With any decision criterion in Table 2, a complete preorder of partial assignments f_A is obtained with respect to mass function m. Therefore, a single but perhaps partial assignment will be selected as the optimal one. Taking complete preference relation $f_{\{\omega_1,\omega_2\}} \succ f_{\omega_1} \succ f_{\omega_2}$ ($|\Omega| = 2$) as an example, we have $\mathscr{F}^* = \{f_{\{\omega_1,\omega_2\}}\}$, i.e., we know for sure that the best choice is to assign the instance to class 1 or class 2, which corresponds to a set-valued prediction $\{\omega_1, \omega_2\}$. We can remark that, in this approach, we can still make set-valued predictions even with very precise probabilities of states of nature, which is a major difference with the other approach discussed in Section 2.1.

Table 1: Decision criteria inducing partial preorders

decision criterion	preference relation
interval dominance	$f_i \succcurlyeq_{ID} f_j \iff \underline{\mathbb{E}}_m(f_i) \geq \overline{\mathbb{E}}_m(f_j)$
maximality	$f_i \succcurlyeq_{max} f_j \iff \underline{\mathbb{E}}_m(f_i - f_j) \geq 0$
weak dominance	$f_i \succcurlyeq_{WD} f_j \iff \left(\underline{\mathbb{E}}_m(f_i) \geq \underline{\mathbb{E}}_m(f_j)\right) \wedge \left(\overline{\mathbb{E}}_m(f_i) \geq \overline{\mathbb{E}}_m(f_j)\right)$

Table 2: Decision criteria inducing complete preorders

decision criterion	preference relation	detail
generalized maximin	$f_{A_i} \succcurlyeq_* f_{A_j} \iff \underline{\mathbb{E}}_m(f_{A_i}) \geq \underline{\mathbb{E}}_m(f_{A_j})$	$\underline{\mathbb{E}}_m(f_{A_i}) = \sum_{B\subseteq\Omega} m(B) \min_{\omega_j\in B} \hat{u}_{A_i,j}$
generalized maximax	$f_{A_i} \succcurlyeq^* f_{A_j} \iff \overline{\mathbb{E}}_m(f_{A_i}) \geq \overline{\mathbb{E}}_m(f_{A_j})$	$\overline{\mathbb{E}}_m(f_{A_i}) = \sum_{B\subseteq\Omega} m(B) \max_{\omega_j\in B} \hat{u}_{A_i,j}$
generalized Hurwicz	$f_{A_i} \succcurlyeq_\alpha f_{A_j} \iff \mathbb{E}_{m,\alpha}(f_{A_i}) \geq \mathbb{E}_{m,\alpha}(f_{A_j})$	$\mathbb{E}_{m,\alpha}(f_{A_i}) = \alpha\underline{\mathbb{E}}_m(f_{A_i}) + (1-\alpha)\overline{\mathbb{E}}_m(f_{A_i})$
pignistic criterion	$f_{A_i} \succcurlyeq_p f_{A_j} \iff \mathbb{E}_p(f_{A_i}) \geq \mathbb{E}_p(f_{A_j})$	$\mathbb{E}_p(f_{A_i}) = \sum_{j=1}^{n} \hat{u}_{A_i,j} BetP(\{\omega_j\})$
generalized OWA	$f_{A_i} \succcurlyeq_\beta f_{A_j} \iff \mathbb{E}^{owa}_{m,\beta}(f_{A_i}) \geq \mathbb{E}^{owa}_{m,\beta}(f_{A_j})$	$\mathbb{E}^{owa}_{m,\beta}(f_{A_i}) = \sum_{B\subseteq\Omega} m(B) F_\beta(\{\hat{u}_{A_i,j} \mid \omega_j \in B\})$
generalized minimax regret	$f_{A_i} \succcurlyeq_r f_{A_j} \iff \overline{R}(f_{A_i}) \leq \overline{R}(f_{A_j})$	$\overline{R}(f_{A_i}) = \sum_{B\subseteq\Omega} m(B) \max_{w_j\in B} r_{A_i,j}$
maximum expected utility	$f_{A_i} \succcurlyeq_m f_{A_j} \iff EU(f_{A_i}) \geq EU(f_{A_j})$	$EU(f_{A_i}) = \sum_{j=1}^{n} \hat{u}_{A_i,j} p_j$

3. Extending Utility Matrix via an OWA Operator

As discussed in Section 2.2, the extended utility matrix $\hat{\mathbb{U}}_{(2^n-1)\times n}$ plays an important role in decision-making. Given original utility matrix $\mathbb{U}_{n\times n}$, we propose to generate $\hat{\mathbb{U}}_{(2^n-1)\times n}$ using an OWA operator.

An OWA operator of dimension n is a mapping F with associated collection of positive weights $\boldsymbol{w} = (w_1, \cdots, w_n)$ summing up to one, such that

$$F(a_1, \cdots, a_n) = \sum_{i=1}^{n} w_i b_i, \tag{1}$$

where b_i is the i-th largest element of $a_1, \cdots, a_n$. By choosing different weights, it provides a parameterized class of mean type aggregation operators. Yager defined the measure of orness [17] as $orness(\boldsymbol{w}) = \frac{1}{n-1}\sum_{i=1}^{n}(n-i)w_i$. Given orness β, the OWA operator F_β that maximizes the entropy $ENT(\boldsymbol{w}) = -\sum_{i=1}^{n} w_i \log w_i$ under the constraint $orness(\boldsymbol{w}) = \beta$ will be chosen.

Now we consider the classification problem. Given a state of nature ω_j, the utility of assigning one instance to set A should intuitively be a function of those utilities of each precise assignments within A. From the most optimistic view, the maximum utility of elements in the set is selected: $\hat{u}_{A,j} = \max_{\omega_i\in A} u_{ij}$. So as long as set A contains the true label, no matter how imprecise A is, the partial assignment achieves utility 1, representing a total tolerance of imprecision. From a more imprecision-neutral point of view, the utility of partial assignment f_A can be defined as the average of the utilities of precise assignments within the set, just as picking one label uniformly at random from set A:

$$\hat{u}_{A,j} = \frac{1}{|A|}\sum_{\omega_i\in A} u_{ij}. \tag{2}$$

We denote the average utility as $\overline{u}_{A,j} := \frac{1}{|A|}\sum_{\omega_i\in A} u_{ij}$. It can be noted that if $\hat{u}_{A,j}$ is less than the average utility $\overline{u}_{A,j}$, it is always preferable to pick a class randomly in A, rather than selecting set A as our prediction. For practical purposes, the utilities of partial assignments can, thus, be defined using a family of parameterized utility functions ranging from the average to the maximum, i.e.,

$$\frac{1}{|A|}\sum_{\omega_i\in A} u_{ij} \leq \hat{u}_{A,j} < \max_{\omega_i\in A} u_{ij},$$

which can be implemented using OWA operators with different weights $\boldsymbol{w}$: given a set $A \subseteq \Omega$ and the state of nature ω_j, the aggregated utility for assigning one instance to set A (denoted as $\hat{u}_{A,j}$) is calculated as a function F of utilities of each elements in this set as

$$\hat{u}_{A,j} = F(\{u_{ij} \mid \omega_i \in A\}) = \sum_{k=1}^{|A|} w_k u^A_{(k)j}, \tag{3}$$

Table 3: Utility matrix extended by an OWA operator with $\gamma = 0.8$

acts	states of nature		
	ω_1	ω_2	ω_3
$f_{\{\omega_1\}}$	1.0000	0.2000	0.1000
$f_{\{\omega_2\}}$	0.2000	1.0000	0.2000
$f_{\{\omega_3\}}$	0.1000	0.2000	1.0000
$f_{\{\omega_1,\omega_2\}}$	0.8400	0.8400	0.1800
$f_{\{\omega_1,\omega_3\}}$	0.8200	0.2000	0.8200
$f_{\{\omega_2,\omega_3\}}$	0.1800	0.8400	0.8400
$f_{\{\omega_1,\omega_2,\omega_3\}}$	0.7373	0.7455	0.7373

where the second equation is the calculation of an OWA operator. $u^A_{(k)j}$ denotes the k-th largest element in the set $\{u_{ij},\ \omega_i \in A\}$, and weight $w_k \geq 0$ represents the DM's preference to choose $u^A_{(k)j}$ when he is forced to make a precise decision among a set of possible choices. Similar to the orness measure, for the OWA operator with weight vector $\boldsymbol{w}$, we define the DM's *tolerance degree of imprecision* as

$$TOL(\boldsymbol{w}) = \sum_{k=1}^{|A|} \frac{|A|-k}{|A|-1} w_k = \gamma, \tag{4}$$

which is equal to 1 for the maximum and 0.5 for the average. Given γ, the weights corresponding to the OWA operator are obtained by maximizing the entropy

$$ENT(\boldsymbol{w}) = -\sum_{k=1}^{|A|} w_k \log w_k, \tag{5}$$

subject to $TOL(\boldsymbol{w}) = \gamma$ and $\sum_{k=1}^{|A|} w_k = 1$.

Table 3 shows the extended utility matrix obtained by an OWA operator with $\gamma = 0.8$, where the first three rows constitute the original utility matrix. Considering the utility matrix shown in Table 3, let us assume that the true label is ω_1. Figure 1 displays the aggregated utilities for sets $\{\omega_1,\omega_2\}$, $\{\omega_1,\omega_3\}$ and $\{\omega_1,\omega_2,\omega_3\}$ with different values of γ. The aggregated utility is only related to the utilities of elements within the set. As γ ranges from 0 to 1, the aggregated utility for each set varies from the minimal utility in this set to the maximal one. When $\gamma = 0.5$, the average utility is obtained by the OWA operator. As mentioned above, we only need to consider values of γ between 0.5 and 1, since when $\gamma < 0.5$ a precise prediction is always more desirable than an imprecise one.

Table 4 shows the utility matrix extended by the "p-discounted cost" approach [20] with $p = 4$. Comparing it to the one extended by an OWA operator (Table 3, $\gamma = 0.8$), we can see the similarity between them. Our method assigns slightly lower utilities to set predictions containing

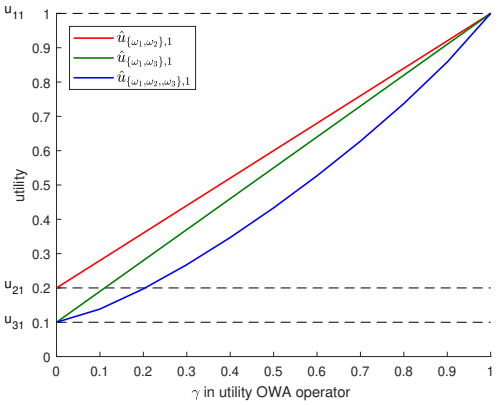

Figure 1: Aggregated utilities vs. tolerance degree of imprecision γ

Table 4: Utility matrix extended by the p-discounted costs approach ($p = 4$)

acts	states of nature		
	ω_1	ω_2	ω_3
$f_{\{\omega_1\}}$	1.0000	0.2000	0.1000
$f_{\{\omega_2\}}$	0.2000	1.0000	0.2000
$f_{\{\omega_3\}}$	0.1000	0.2000	1.0000
$f_{\{\omega_1,\omega_2\}}$	0.8412	0.8412	0.1707
$f_{\{\omega_1,\omega_3\}}$	0.8409	0.2000	0.8409
$f_{\{\omega_2,\omega_3\}}$	0.1707	0.8412	0.8412
$f_{\{\omega_1,\omega_2,\omega_3\}}$	0.7602	0.7604	0.7602

the true label and slightly higher utilities to imprecise and incorrect predictions. We can also briefly check the properties of utility generated by an OWA operator with respect to the guidelines proposed by Yang et al. [20]. As $\gamma = 0.5$ corresponds to the average utility $\overline{u}_{A,j}$, for any $\gamma \in (0.5, 1)$, the necessary properties (Properties 1, 3, 4 and 10 in [20]) and desirable properties (Properties 2 and 5) are satisfied. Regarding context-dependent properties, for any set-valued prediction A, we have $\omega_j \notin A \Rightarrow u^{\gamma}_{A,j} \geq u^{\gamma=0.5}_{A,j} = \overline{u}_{A,j}$, satisfying Property 7. Since different utilities can be assigned in $\mathbb{U}$ according to the truth, our proposal satisfies Property 9.

4. Evaluation of Set-valued Predictions

In classification applications, a test set T is used to assess the performance of the learned model. When we make set-valued predictions according to strategies described in Section 2, a standard is needed to evaluate the decisions made by different criteria.

Zaffalon [21] proposed a utility-discounted predictive accuracy under the $\{0,1\}$ reward assumption to evaluate set-valued predictions made of credal classifiers. In this paper, we propose to use the extended utility matrix $\hat{\mathbb{U}}$ generated by an OWA operator for performance evaluation. The classification performance is evaluated by the *averaged utility* in the test set T:

$$Acc(T) = \frac{1}{|T|}\sum_{i=1}^{|T|} \hat{u}_{\mathscr{F}_i^*,i^*}, \tag{6}$$

where $\hat{u}_{\mathscr{F}_i^*,i^*}$ denotes the utility of selecting the optimal act $f_{\mathscr{F}_i^*}$ (assigning instance i in T to set $\mathscr{F}_i^*$) when its true class is ω_i^*.

5. Experiments

In this section, we report on classification experiments aiming to compare experimentally the different decision strategies described above. Belief functions concerning the states of nature were generated through the DS theory-based neural network classifier [5], which assigns for each instance a mass to each singleton class and the frame of discernment, i.e.,

$$m(\omega_i) = m_i,\ i = 1,\ldots,n;\ m(\Omega) = 1 - \sum_{i=1}^{n} m_i.$$

5.1. Classification Performances with Varying γ

We first checked the averaged utilities obtained according to different decision criteria with varying γ. Experiments were carried out using the UCI Balance scale dataset [12], which is a four-attribute and three-class dataset containing 625 instances. The original utility matrix was arbitrarily assumed to be

$$\mathbb{U} = \begin{bmatrix} 1 & 0.2 & 0.1 \\ 0.2 & 1 & 0.2 \\ 0.1 & 0.2 & 1 \end{bmatrix},$$

where correct predictions (diagonal elements) had utilities 1 and different prediction errors were assumed to be not treated equivalently. This is just an example to show the generality of the approach; in most cases, matrix $\mathbb{U}$ can be taken to be the identity matrix, in which only correct prediction can achieve utility 1 while all the other outcomes achieve 0. To evaluate the results, five-fold cross-validation was performed, and all experiments were repeated five times to compute an average result.

Table 5 shows the averaged utilities (upper part) and corresponding percentage of precise prediction (lower part) in each case. Given a fixed γ, the averaged utilities of different criteria vary in a narrow range (the maximum averaged utilities are highlighted in bold). For criteria with complete preorders among partial assignments (DC1-DC6), as γ increases, imprecise predictions are more preferred so the percentage of precise predictions decreases; the averaged utilities decrease slightly and then increase to 1. To better explain this behaviour, Table 6 reports the predictions and corresponding utilities of three instances, where ω^* is the true label, and $\Omega = \{\omega_1, \omega_2, \omega_3\}$. For misclassified instance (#2), its utility can increase from 0.2 to 1 as γ becomes larger. Yet for those correctly predicted when $\gamma = 0.5$ (#1 and #3), their utilities will drop and then increase back to 1. The majority of instances are of the latter case, so overall, the averaged utilities decrease at first as set-valued predictions containing the true label have lower utility than the precise ones. As γ approaches 1, the utilities of imprecise predictions grow closer to 1, making the averaged utility increase to 1.

For decision criteria yielding partial preorders among precise assignments (DC7-DC9), the extended utility matrix is only used for evaluation, so the predictions remain unchanged as γ increases. The weak dominance criterion (DC9) nearly always gives precise predictions (whose utilities are not affected by γ) for this dataset; consequently, averaged utilities do not change with γ. The averaged utilities of the other two criteria (DC7 and DC8) increase monotonically when γ increases from 0.5 to 1. The extended utility matrix gives the same set-valued prediction a higher utility as γ grows. So even though the predictions themselves are not affected by γ, they do achieve higher averaged utilities with larger γ.

5.2. Performances with Noised Test Sets

In many situations, a classifier is trained with "good" data (acquired and preprocessed in controlled conditions) and then used in a real environment where, for instance, sensors may be not well calibrated. In such a case, the test data do not have the same distribution as the learning data. Cautious decision rules making set-valued predictions can be expected to be particularly beneficial in such an environment, and discrepancies between the performances of different decisions rules may be more apparent than they are in the case of "clean" data considered in Section 5.1.

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 sets were simulated from three Gaussian distributions with the following characteristics:

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

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

where I is the identity matrix. Figure 2 visualizes a particular dataset of 600 instances generated in this way.

To simulate a different distribution, a random noise was added to the features of the test instances by Algorithm 1.

Table 5: Results for the Balance scale data set with different values of γ

		DC1	DC2	DC3	DC4	DC5	DC6	DC7	DC8	DC9
averaged utility	γ=0.5	0.9186	**0.9188**	0.9186	0.9186	0.9186	0.9186	0.9187	0.9187	0.9187
	γ=0.6	0.9179	0.9184	0.9176	0.9179	0.9184	0.9176	**0.9188**	**0.9188**	0.9187
	γ=0.7	0.9059	0.9064	0.9052	0.9059	0.9056	0.9054	**0.9190**	**0.9190**	0.9187
	γ=0.8	0.9043	0.9032	0.9028	0.9043	0.9030	0.9024	**0.9191**	**0.9191**	0.9188
	γ=0.9	**0.9339**	0.9319	0.9325	**0.9339**	0.9331	0.9319	0.9192	0.9192	0.9188
	γ=1.0	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.9194	0.9194	0.9188
% of precision	γ=0.5	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	97.44%	97.44%	99.97%
	γ=0.6	88.96%	89.47%	88.96%	88.96%	89.18%	89.06%	97.44%	97.44%	99.97%
	γ=0.7	80.10%	80.77%	80.06%	80.10%	80.22%	80.26%	97.44%	97.44%	99.97%
	γ=0.8	69.70%	70.14%	69.63%	69.70%	69.82%	69.63%	97.44%	97.44%	99.97%
	γ=0.9	57.02%	57.76%	57.12%	57.02%	57.38%	57.12%	97.44%	97.44%	99.97%
	γ=1.0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	97.44%	97.44%	99.97%

[1] DC1: maximin DC2: maximax DC3: pignistic DC4: minimax regret DC5: Hurwicz ($\alpha = 0.6$) DC6: OWA ($\beta = 0.6$) DC7: interval dominance DC8: maximality DC9: weak dominance

Table 6: Label predictions/utilities with different γ

γ	#1 ($\omega^* = \omega_3$)	#2 ($\omega^* = \omega_2$)	#3 ($\omega^* = \omega_3$)
0.5	ω_3/1	ω_1/0.2	ω_3/1
0.6	ω_3/1	ω_1/0.2	$\{\omega_1, \omega_3\}$/0.64
0.7	$\{\omega_1, \omega_3\}$/0.73	ω_1/0.2	$\{\omega_1, \omega_3\}$/0.73
0.8	$\{\omega_1, \omega_3\}$/0.82	$\{\omega_1, \omega_2\}$/0.84	$\{\omega_1, \omega_3\}$/0.82
0.9	$\{\omega_1, \omega_3\}$/0.91	Ω/0.8610	Ω/0.8584
1.0	Ω/1	Ω/1	Ω/1

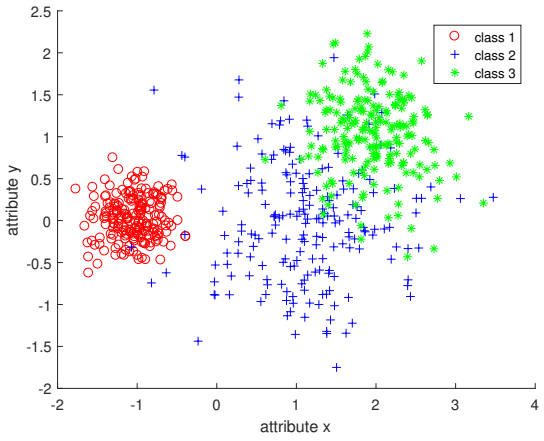

Figure 2: A Gaussian data set of 600 instances

Algorithm 1: Algorithm to generate a noisy test set

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

We set $\gamma = 0.8$ and let the noise standard deviation σ vary from 0 to 10 to simulate different levels of noise. The experiments were repeated 20 times to compute an average result. In each experiment, the training and test sets contained 600 and 300 instances, respectively.

With higher noise level, the distribution of the test set becomes more different from that of the training set. The averaged utilities are plotted against the noise level according to various decision criteria in Figure 3. Similar to the previous experiment, the percentage of precise predictions shown in Figure 4 helps to analyse the performances.

For $\sigma = 0$, the test set and the training set have the same distribution; the criteria based on partial preorder achieve lower averaged utilities as they make more precise but incorrect predictions. When σ increases, the averaged utilities for all criteria drop quickly and the performances of different criteria start to differ. When the test set distribution becomes more different, the maximax and weak dominance criteria perform worse than others, as they make precise predictions most of the time. For the other seven criteria, as σ varies from 3 to 10, the averaged utilities remain stable or even increase slightly. Basically, when uncertainty increases, the decision criteria (except maximax

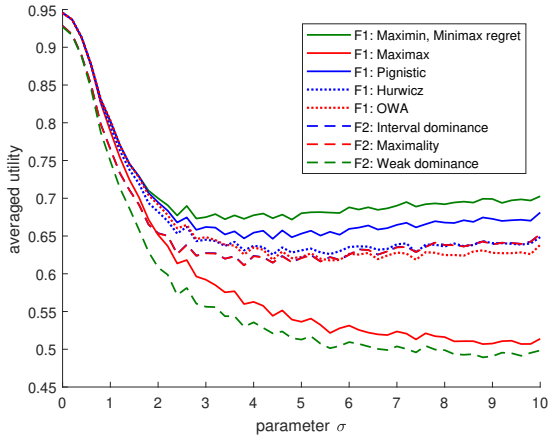

Figure 3: Averaged utilities of different criteria as a function of noise level

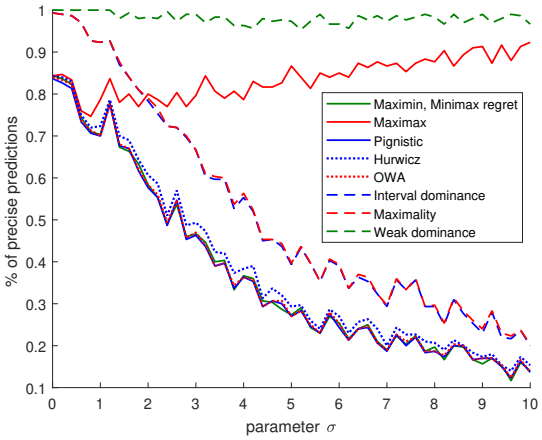

Figure 4: Percentage of precise predictions of different criteria as a function of noise level

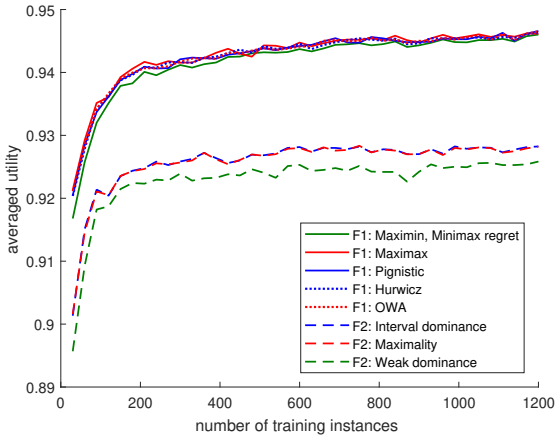

Figure 5: Averaged utilities of different criteria as a function of training set size ($\alpha_m = 0.99$)

and weak dominance) assign more instances to sets. As different classes overlap more with larger σ, imprecise predictions are more likely to contain the true labels, achieving a higher averaged utility. The maximin criterion is the most conservative, resulting in fewer misclassifications.

5.3. Performances with Increasing Training Set Size

In the third experiment, we compared the performances of different decision criteria as the size of the training set increases. We kept the same experimental settings as in Section 5.2. The size of the training set was increased from 60 to 1200. Twenty training sets of each size were randomly generated and the average result was considered. A test set of 300 instances was used for performance evaluation. Figures 5 and 6 display the averaged utilities for different decision criteria for two settings of the DS neural network classifier[1], respectively, $\alpha_m = 0.99$ (more specific belief functions) and $\alpha_m = 0.8$ (less specific belief functions). In Figures 7 and 8, we also give the percentages of precise predictions for, respectively, $\alpha_m = 0.99$ and $\alpha_m = 0.8$.

When $\alpha_m = 0.99$ (Figures 5 and 7), the output belief functions are close to probabilities. Overall, all criteria based on a complete preorder of partial assignments perform similarly, and significantly better than the criteria based on a partial preorder. When the training set size is smaller than 240, the increase of averaged utility is mainly due to the increasing proportion of precise and correct

1. Parameter $0 < \alpha_m < 1$ controls the masses given to Ω and each singleton class ω_i in the DS neural network classifier (Eqs (17) and (18) in Ref. [5]). The belief functions provided by the classifier become more specific as α_m increases, say, $m(\omega_1) = 0.02$, $m(\omega_2) = 0.97$, $m(\Omega) = 0.01$ when $\alpha_m = 0.99$ and $m(\omega_1) = 0.07$, $m(\omega_2) = 0.90$, $m(\Omega) = 0.03$ when $\alpha_m = 0.8$.

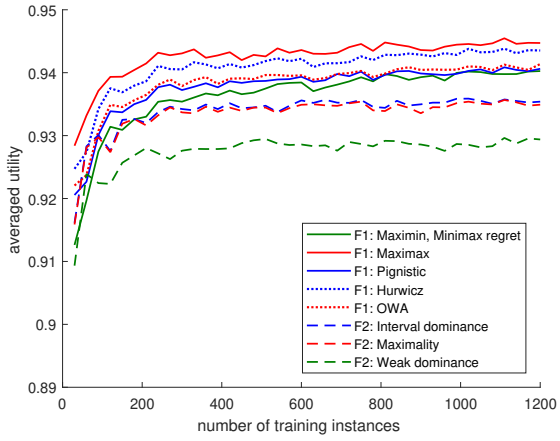

Figure 6: Averaged utilities of different criteria as a function of training set size ($\alpha_m = 0.8$)

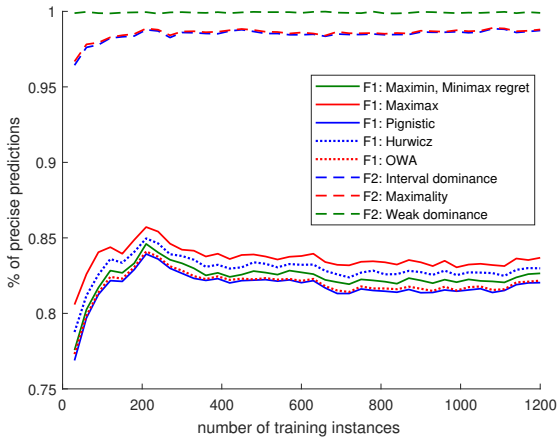

Figure 7: Percentage of precise predictions as a function of training set size ($\alpha_m = 0.99$)

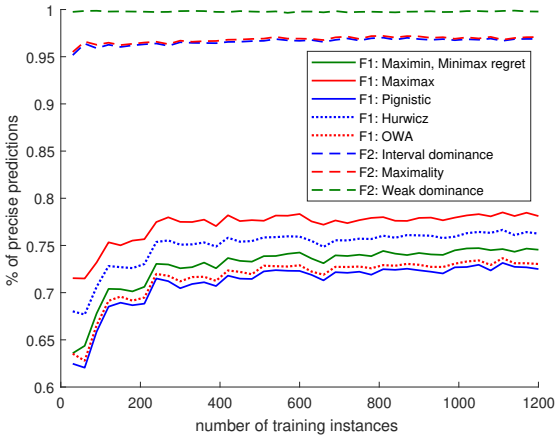

Figure 8: Percentage of precise predictions as a function of training set size ($\alpha_m = 0.8$)

predictions. When the size continues to grow, more imprecise predictions are made. The averaged utilities increase slightly as some misclassified instances become imprecisely but correctly predicted ones. Compared to the OWA decision criterion, the maximax criterion makes more precise predictions but has similar averaged utility.

When the belief functions are less specific (Figures 6 and 8), as the size of the training set increases, more precise predictions lead to an increase in averaged utilities. The maximax and Hurwicz criteria work best in this case. The maximin criterion, which is the most conservative, has relatively low averaged utilities as its predicted sets are larger than others.

For both settings $\alpha_m = 0.99$ and $\alpha_m = 0.8$, the weak dominance criterion gives precise predictions almost all the time and has the worst performance. All the other criteria lead to set-valued predictions, which results in higher averaged utilities.

In addition, it is also notable that for all the experiments, the interval dominance and maximality criteria have quite similar performances. The optimal set of maximality $\mathscr{F}^*_{max}$ is included in that of interval dominance $\mathscr{F}^*_{ID}$. For the data sets used in this paper, $\mathscr{F}^*_{ID}$ yields singleton predictions most of the times, leaving little space for maximality to provide different decisions.

6. Conclusion

To make set-valued predictions in evidential classification problems, decision criteria can be based either on a partial preorder among precise assignments, or on a complete preorder among partial assignments. Using an extended utility matrix generated via an OWA operator, experimental com-

parisons were performed on UCI and simulated Gaussian data sets. The set-valued predictions induced by a partial preorder turn into precise ones when information becomes more precise. In contrast, the criteria based on a complete preorder can provide set-valued predictions even when uncertainty is quantified by probabilities. Experimental results suggest that set-valued predictions perform better than precise ones in the case of complex data sets: therefore, the most cautious rules should be preferred in highly uncertain environments. More experiments are needed to determine which decision rules should be recommended for different classification problems.

Whereas the Dempster-Shafer setting was assumed in this paper, a similar analysis could be carried out in other settings such as the imprecise probability framework [1]. Also, the belief functions used in this paper were obtained with a particular evidential classifier. In future work, we will consider other DS theory-based classifiers such as described in [10, 6], and we will analyse the performances of various decision rules when mass functions have more general focal sets. The various approaches to decision-making in evidential classifications will also be explored more thoroughly, both theoretically and empirically.

Acknowledgments

This research was supported by Shandong Provincial Natural Science Foundation ZR2018PF009, Shandong Provincial Key Research and Development Program 2017GGX10116 and the China Scholarship Council. It was also supported by the Labex MS2T funded by the French Government through the program "Investments for the future" by the National Agency for Research (reference ANR-11-IDEX-0004-02).

References

[1] Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[2] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, pages 325–339, 1967.

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

[4] Thierry Denoeux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

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

[6] Thierry Denœux. Logistic regression, neural networks and Dempster-Shafer theory: A new perspective. *Knowledge-Based Systems*, 2019. doi: https://doi.org/10.1016/j.knosys.2019.03.030.

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

[8] Thierry Denœux, Didier Dubois, and Henri Prade. Representations of uncertainty in artificial intelligence: Beyond probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, chapter 4. Springer Verlag, 2019.

[9] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[10] Orakanya Kanjanatarakul, Siwarat Kuson, and Thierry Denœux. An evidential k-nearest neighbor classifier based on contextual discounting. In Fabio Cuzzolin, Thierry Denœux, Sébastien Destercke, and Arnaud Martin, editors, *Belief Functions: Theory and Applications: Fourth International Conference (BELIEF 2018)*, number 11069 in Lecture Notes in Artificial Intelligence, pages 155–162. Springer, Compiègne, France, Sept. 2018.

[11] Chunfeng Lian, Su Ruan, and Thierry Denœux. An evidential classifier based on feature selection and two-step classification strategy. *Pattern Recognition*, 48(7):2318–2327, 2015.

[12] M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[13] Zhun-Ga Liu, Quan Pan, Jean Dezert, and Arnaud Martin. Combination of classifiers with optimal weight based on evidential reasoning. *IEEE Transactions on Fuzzy Systems*, 26(3):1217–1230, 2018.

[14] Glenn Shafer. *A mathematical theory of evidence*, volume 1. Princeton University Press, Princeton, 1976.

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

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

[17] Ronald R Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on systems, Man, and Cybernetics*, 18(1):183–190, 1988.

[18] Ronald R. Yager. Decision making under Dempster-Shafer uncertainties. *International Journal of General Systems*, 20(3):233–245, 1992.

[19] Ronald R. Yager. Decision making using minimization of regret. *International Journal of Approximate Reasoning*, 36(2):109–128, 2004.

[20] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. The costs of indeterminacy: How to determine them? *IEEE transactions on cybernetics*, 47 (12):4316–4327, 2017.

[21] Marco Zaffalon, Giorgio Corani, and Denis Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282, 2012.