16th World Congress of the International Fuzzy Systems Association (IFSA)
9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)

Generalized stochastic orderings applied to the study of performance of machine learning algorithms for low quality data

Inés Couso[1] **Luciano Sánchez**[2]

[1]Department of Statistics, Universidad de Oviedo, Spain
[2]Department of Computer Sciences, Universidad de Oviedo, Spain

Abstract

Usually, the expected loss minimization criterion is used in order to look for the optimal model that expresses a certain response variable as a function of a collection of attributes. We generalize this criterion, in order to be able to deal also with those situations where a numerical loss function makes no sense or is not provided by the expert. In a first stage, we consider the new framework in standard situations, where both the collection of attributes and the response variables are observed with precision. In a second one, we assume that we are just provided with imprecise information about them (in terms of set-valued data sets). We cast some comparison criteria from the recent literature on learning methods from low-quality data as particular cases of our general approach.

Keywords.- Regression, classification, loss function, generalized stochastic ordering, set-valued data, low-quality data.

1. Introduction

We deal with those machine learning problems where the goal is to find an optimal model $f : \mathcal{X} \to \mathcal{Y}$ relating some response variable $Y : \Omega \to \mathcal{Y}$ to a collection of attributes $\mathbf{X} : \Omega \to \mathcal{X}$, both of them defined on the same population Ω. This kind of optimization problems usually aim at minimizing the expected loss, according to some loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that assigns a specific value to every pair $(Y(\omega), f(\mathbf{X}(\omega)))$, composed by the outcome of the response variable and its estimate based on the collection of attributes, for every individual $\omega \in \Omega$. Very typical examples of that are the square and the absolute value of the difference ($\Delta(y, \hat{y}) = (y - \hat{y})^2$) and $\Delta'(y, \hat{y}) = |y - \hat{y}|$, respectively), both of them commonly used in regression problems, as well as the $0-1-$valued loss function $\Delta(y, \hat{y}) = 1_{\hat{y} \neq y}$, frequently used in classification problems. But sometimes, a numerical valued loss function is impossible to assess. For instance, an expert can tell us that classifying a girl with severe dyslexia as non-dyslexic is worse than classifying her as having a moderate dyslexia, but he may be unable to provide us with specific loss values on a numerical scale.

On the other hand during the last years, there has been a growing interest ([7, 8, 10, 15, 22, 24]) in the development of learning models from set-valued datasets, extending existing learning algorithms to the case where our data points are not elements in the cartesian product $\mathcal{X} \times \mathcal{Y}$ anymore, but (crisp or fuzzy) subsets of it. In order to do so, one should first distinguish between the "ontic" and the "epistemic" interpretations of set-valued data (see [3, 11]). Under the ontic approach (also called the "conjunctive" approach), sets are understood as complex entities observed with precision. As pointed out by E. Hüllermeier in [15], this interpretation suggests learning models that produce sets as predictions, i.e., models that *reproduce* the observed data. Thus, methods based on this interpretation of (fuzzy) sets usually produce parametric models where the parameters are indeed subsets of the parametric space, instead of elements of it. On the contrary, under the epistemic approach (also called the "disjunctive" approach), sets are used to describe our (in)complete knowledge about the true outcomes of the vector of attributes and/or the response variable: we do not observe their exact values, but we just can provide sets that contain them with total certainty. In that case, we aim to find a crisp model that relates the (possibly ill-observed) response variable to the (also possibly ill-observed) random vector of attributes. Models are therefore usual functions of the form $f : \mathcal{X} \to \mathcal{Y}$, but our information about their respective performances over a particular individual ω is incomplete, and it can be naturally expressed in terms of a subset of the form:

$$\{\Delta(y, f(\mathbf{x})) \, : \, \mathbf{x} \in \mathbb{X}(\omega), y \in \mathbb{Y}(\omega)\},$$

where $\mathbb{X}(\omega)$ and $\mathbb{Y}(\omega)$ denote the most precise sets that respectively contain $\mathbf{X}(\omega)$ and $Y(\omega)$ with certainty, according to our incomplete information. Our information about the expected loss is therefore also incomplete, and a partial or a total ordering over the class of subsets of the real line, extending the usual order, needs to be considered, in order to compare two different models. This technique gives birth to different extended methods, depending on the nature of the algorithms to be extended and the partial/total ordering selected ([17, 18, 19, 20, 24, 25, 26]). This paper makes use

of these techniques in a more general context, where a numerical loss function is not necessarily defined, and therefore an expected loss minimization does not necessarily make sense. In our framework, every model is identified with a random variable representing its reward (the opposite to its loss). According to this view, any pairwise comparison criterion between models can be identified with a stochastic ordering. In this context, the expected loss minimization criterion is based on a well known stochastic ordering called "dominance in the sense of expected utility". Such a particular stochastic ordering criterion is replaced by a general family of criteria involving a wider family of stochastic orderings, where the "expectation dominance" criterion is just a particular example. Other stochastic orderings do not require the compared random elements to be numerical valued, and therefore non numeric loss functions can be considered in our general environment.

Our goal in the paper is not to argue that any particular method for the selection of optimal (or, more generally, non-dominated) models is better than another, but simply to present a general framework that facilitates the study of formal relations between different selection criteria.

2. Basic notions and nomenclature

Let Ω denote the population under study. Let $\mathcal{X}$ be the input space (the set of possible outcomes of the vector of attributes) and let $\mathcal{Y}$ denote the output space, i.e, the set of possible outcomes for the response variable. For instance, in a regression problem, the output space may coincide with the real line, while in a classification problem it will represent the collection of classes. $\mathbf{X} = (X_1, \ldots, X_d) : \Omega \to \mathcal{X}$ will denote the random vector of attributes and $Y : \Omega \to \mathcal{Y}$ will stand for the response variable (in particular, in classification problems, $Y(\omega)$ will represent the class of object ω.)

We will consider a loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, penalizing wrong predictions. A model will incur a penalty $\Delta(y, \hat{y})$ if the true output value is y and the model predicts $\hat{y}$. Examples of loss functions commonly used in regression problems are the so-called *square loss* $\Delta(y, \hat{y}) = (y - \hat{y})^2$, the *absolute loss*, $\Delta(y, \hat{y}) = |y - \hat{y}|$, and the ϵ-*insensitive loss* $\Delta(y, \hat{y}) = (|y - \hat{y}| - \epsilon) \cdot 1_{|y-\hat{y}|>\epsilon}$. In classification problems, the *zero-one loss* function $\Delta(y, \hat{y}) = 1_{y \neq \hat{y}}$ is probably one of the most common loss function used in the literature. Other non-symmetrical cost functions determined by means of a $k \times k$ cost matrix will be also mentioned in this paper.

We will call the *risk* associated to a *model* $f : \mathcal{X} \to \mathcal{Y}$ to the expected loss:

$$E\left[\Delta(Y, f(\mathbf{X}))\right] = \int_{\Omega} \Delta(Y(\omega), f(\mathbf{X}(\omega)))\, dP(\omega),$$

where P is a probability measure defined on the population Ω and inducing a joint probability distribution $P_{(\mathbf{X},Y)}$ on $\mathcal{X} \times \mathcal{Y}$. We assume that P is unknown, and therefore we cannot assess in practice the expected reward to a particular model f. Notwithstanding, nothing prevents us from referring to such an expected reward from a theoretical point of view. It is usual to estimate the risk of a model on the basis of a sample of observations. For instance, we can take a finite set of pairs $\mathcal{S} = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ representing a realization of a sequence of i.i.d. vectors following the distribution $P_{(\mathbf{X},Y)}$. The performance of a model f on the sample $\mathcal{S}$ is described by the empirical risk

$$\mathcal{R}^{\mathcal{S}}_{\Delta,f} = \frac{1}{n} \sum_{i=1}^{n} \Delta(y_i, f(\mathbf{x}_i)).$$

A review of existing methods of estimation of the performance of models, as well as the proposal of new ones is out of the scope of this paper.

In this paper, we will consider a more general setting were the loss function Δ does not necessarily take values on the real line, but on a partial (pre)ordered set, $(\mathcal{O}, \leq_{\mathcal{O}})$ (a *preposet*). With this generalized setting, we will be able to address those situations where the expert cannot assess a numerical penalty value to each wrong prediction, but (s)he can establish a (possibly partial) ordering between different (wrong) predictions. For instance, a medical doctor may tell us that assuming that patients are sick, when they are in fact healthy is worse than doing the contrary, but without assessing a numerical penalty value to any of those misclassifications. This framework contains, among others, the traditional above mentioned cases (square loss, absolute loss, $\epsilon-$insensitive loss, zero-one loss), as well as non-symmetric loss functions arising in imbalanced classification problems. In this general context, we cannot make use of the risk function, since it is calculated as a Lebesgue expectation and therefore it requires the existence of a numerical-valued loss function. The following example illustrates how arbitrary choices of loss functions can lead us to completely different comparisons.

Example 1 *Consider a binary classification problem, where the classes are "healthy" and "sick", for a certain illness. The expert informs us that assuming that patients are sick, when they are in fact healthy is worse than doing the contrary. Mathematically,*

$$\Delta(\mathit{healthy}, \mathit{sick}) > \Delta(\mathit{sick}, \mathit{healthy}).$$

We are not provided with specific penalty values, but, if we want to calculate the expected loss associated to a classifier, we need to define them. The following cost matrices, respectively associated to loss functions Δ_1 and Δ_2 agree with the information given by the expert:

$Y \backslash f(\mathbf{X})$	*healthy*	*sick*
healthy	*0*	*1*
sick	*0.6*	*0*

$Y \backslash f(\mathbf{X})$	*healthy*	*sick*
healthy	*0*	*1*
sick	*0.4*	*0*

The following joint frequency tables display the outcomes of two different classifiers in a sample of 100 patients, where 77 of them are healthy and 23 of them are sick.

Classifier 1:

$Y \backslash f_1(\mathbf{X})$	*healthy*	*sick*
healthy	*70*	*7*
sick	*3*	*20*

Classifier 2:

$Y \backslash f_2(\mathbf{X})$	*healthy*	*sick*
healthy	*69*	*8*
sick	*1*	*22*

The calculation of the empirical risk reports the following results:

- $\mathcal{R}^{\mathcal{S}}_{\Delta_1,f_1} = \frac{7\cdot 1+3\cdot 0.6}{100} = 0.088.$
- $\mathcal{R}^{\mathcal{S}}_{\Delta_1,f_2} = \frac{8\cdot 1+1\cdot 0.6}{100} = 0.086.$
- $\mathcal{R}^{\mathcal{S}}_{\Delta_2,f_1} = \frac{7\cdot 1+3\cdot 0.4}{100} = 0.082.$
- $\mathcal{R}^{\mathcal{S}}_{\Delta_2,f_2} = \frac{8\cdot 1+1\cdot 0.4}{100} = 0.084.$

According to the above calculations, Model 2 performs better than Model 1, if we choose the first loss function, Δ_1, to determine their respective assessments. But we obtain the opposite result if we use the second loss function, Δ_2, to compare them. Different numerical penalty choices for the misclassification of a sick patient lead us to different results about the performances of both models.

The next section presents a general framework where different alternatives to the risk function can be considered.

3. Stochastic orderings in machine learning problems

As we pointed out in last section, we will focus on those problems aiming to find the best model (or a set of non-dominated ones) with respect to the risk function. Our objective function can be expressed either as a loss function or as its opposite (sometimes called a *reward function* or a *utility function*), in which case it is to be maximized. In the rest of the paper, we will consider this dual setting, in order to match the usual literature on stochastic orderings. We will call the *reward* associated to a model f to the random variable $U^{(\mathbf{X},Y)}_{\Delta,f} : \Omega \to \mathbb{R}$ defined as follows:

$$U^{(\mathbf{X},Y)}_{\Delta,f}(\omega) = -\Delta(Y(\omega), f(\mathbf{X}(\omega))),\ \forall\, \omega \in \Omega. \quad (1)$$

According to this dual formulation, we aim to find a model f that maximizes the expected reward, with respect to the probability measure P on Ω, $E[U^{(\mathbf{X},Y)}_{\Delta,f}] = \int U^{(\mathbf{X},Y)}_{\Delta,f}(\omega)\, dP(\omega)$.

With this nomenclature in mind, we identify every model f with its respective reward function $U^{(\mathbf{X},Y)}_{\Delta,f}$. According to this identification, a model f is said to be "better" than another model f', when the expectations of their respective reward functions, $E[U^{(\mathbf{X},Y)}_{\Delta,f}]$ and $E[U^{(\mathbf{X},Y)}_{\Delta,f}]$, satisfy the inequality $E[U^{(\mathbf{X},Y)}_{\Delta,f}] \geq E[U^{(\mathbf{X},Y)}_{\Delta,f}]$. This kind of comparison between both reward functions reminds us to a particular stochastic ordering criterion in Probability Theory called "dominance in the sense of expected utility" ([21]). Stochastic ordering criteria aim to establish preference orderings between random variables defined on a probability space. According to "dominance in the sense of expected utility", a random variable $Z : \Omega \to \mathbb{R}$ is said to be preferred to another one $Z' : \Omega \to \mathbb{R}$ whenever $E[Z] = \int Z(\omega)\, dP(\omega)$ is greater than or equal to $E[Z'] = \int Z'(\omega)\, dP(\omega)$. But as we have clarified in Section 2, sometimes the expert assessments cannot be represented by means of a numerical loss/reward function. From now on, given a particular (generalized) loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathcal{O}$, we will identify any model $f : \mathcal{X} \to \mathcal{Y}$ with its *(generalized) reward* $U^{(\mathbf{X},Y)}_{\Delta,f} : \Omega \to \mathcal{O}$. I.e., we will identify any model $f : \mathcal{X} \to \mathcal{Y}$ with a mapping $U^{(\mathbf{X},Y)}_{\Delta,f}$, defined on Ω and taking values on the preposet, $(\mathcal{O}, \leq_{\mathcal{O}})$. When the expert provided us with a numerical loss/reward function, we represented the performance of a model in terms of its expected reward. Here we aim to propose a general performance assessment procedure that includes, as a particular case, the expected reward approach, but also encompasses other performance assessments that can be used in those situations where non numerical reward functions are provided. In this general setting, we can establish pairwise comparisons between models, without previously establishing performance degrees to single models. A general comparison rule can be established as follows.

Definition 1 *Let Ω be the initial space (space of states of nature or population) where the vector of attributes $\mathbf{X}$ and the response variable Y are defined. Let $\mathcal{X}$ and $\mathcal{Y}$ respectively denote the input and the output spaces. Let us consider a partial preordering, $\preceq$ defined on the family of random mappings from Ω to a preposet $\mathcal{O}$. Let $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathcal{O}$ denote a loss function on $\mathcal{Y}$. A model $f : \mathcal{X} \to \mathcal{Y}$ is said to be $\preceq -\Delta-$*better *than another one $f' : \mathcal{X} \to \mathcal{Y}$ whenever $U^{(\mathbf{X},Y)}_{\Delta,f'} \preceq U^{(\mathbf{X},Y)}_{\Delta,f}$.*

According to the last definition, and on the basis of a particular (generalized) loss function: $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathcal{O}$, we propose to base our comparison between models f and f' on some specific stochastic criterion involving their respective rewards, $U^{(\mathbf{X},Y)}_{\Delta,f} : \Omega \to \mathcal{O}$ and $U^{(\mathbf{X},Y)}_{\Delta,f'} : \Omega \to \mathcal{O}$, not necessarily the expec-

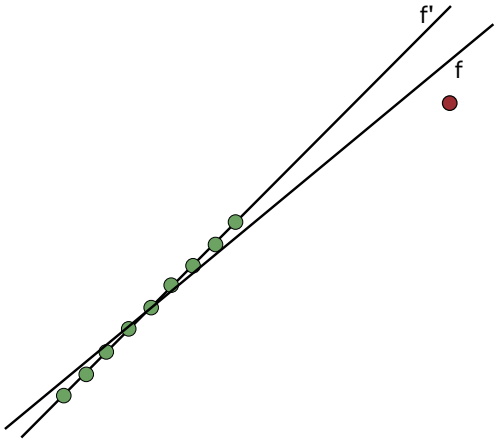

Figure 1: Mean square error vs statistical preference wrt square loss.

tation dominance criterion. Below, we adapt the above general rule to the particular notions of "statistical preference" ([6]) and "first stochastic dominance" ([14]), which are the most common stochastic orderings in classical Probability Theory literature, along with the criterion of dominance in the sense of expected utility above mentioned. Unlike the expected utility criterion, these last two criteria do not require the reward function to be numerical valued.

Definition 2 *Given two models $f : \mathcal{X} \to \mathcal{Y}$ and $f' : \mathcal{X} \to \mathcal{Y}$, $f \preceq_{1st} -\Delta-$ dominates f' if*

$$P(U^{(\mathbf{X},Y)}_{\Delta,f} >_{\mathcal{O}} c) \geq P(U^{(\mathbf{X},Y)}_{\Delta,f'} >_{\mathcal{O}} c),\ \forall\, c \in \mathcal{O}.$$

Definition 3 *Given two models $f : \mathcal{X} \to \mathcal{Y}$ and $f' : \mathcal{X} \to \mathcal{Y}$, $f \preceq_{sp} -\Delta-$ dominates f' if*

$$P(U^{(\mathbf{X},Y)}_{\Delta,f} >_{\mathcal{O}} U^{(\mathbf{X},Y)}_{\Delta,f'}) \geq P(U^{(\mathbf{X},Y)}_{\Delta,f'} >_{\mathcal{O}} U^{(\mathbf{X},Y)}_{\Delta,f}).$$

In the following examples, we compare pairs of models in some regression and classification toy problems using some of the above mentioned criteria.

Example 2 *In this example we will illustrate the pairwise comparison based on statistical preference (Definition 3) against the usual expectation dominance criterion. Let us consider a simple regression problem, involving two variables X and Y observed in a sample of size $n = 10$. Figure 1 displays the dataset along with a pair of linear models f and f'.*

Let us consider the square loss, $\Delta(y, \hat{y}) = (y - \hat{y})^2,\ \forall\, y, \hat{y} \in \mathbb{R}$. According to the criterion of expectation dominance, f is preferred to f'. In fact, f is the regression line (it maximizes the expected empirical reward, mathematically: $f = \arg\max_g \frac{1}{10}\sum_{i=1}^{10} U^{(X,Y)}_{\Delta,g}(\omega_i) = \arg\min \frac{1}{10}\sum_{i=1}^{10}[y_i - g(x_i)]^2$). On the contrary, according to the criterion of statistical preference, f' is preferred to f. The estimation of the probabilities $P(U^{(\mathbf{X},Y)}_{\Delta,f} > U^{(\mathbf{X},Y)}_{\Delta,f'})$ and $P(U^{(\mathbf{X},Y)}_{\Delta,f'} > U^{(\mathbf{X},Y)}_{\Delta,f})$ in our sample, will be given by the respective relative frequencies:

$$\frac{1}{10}\#\{i \in \{1,\ldots,10\} : |y_i - f(x_i)| < |y_i - f'(x_i)|\} \tag{2}$$

$$\frac{1}{10}\#\{i \in \{1,\ldots,10\} : |y_i - f(x_i)| > |y_i - f'(x_i)|\} \tag{3}$$

We easily deduce from Figure 1 that the relative frequency calculated in Equation 3 is less than the one calculated in Equation 2. In fact, this second criterion (statistical preference combined with the -squared- error) is based on the number of individuals in the sample for which the estimation $f(x_i)$ is closer to y_i than the estimation $f'(x_i)$ is, but without taking into account the magnitude of those distances. We have colored the points in the scatter plot green or red, depending on their relative vertical distances to both lines. There are nine green points against just one red one.

Example 3 *We will illustrate here the criterion proposed in Definition 3 (the one based on first stochastic dominance) against the usual criterion based on the expectation of rewards. Let us consider the sample of Example 2. Figure 2 displays the scatter plot along with the pair of models f and f'. For a fixed distance $\epsilon > 0$, and for each model g, we consider the number of instances where the distance between response value y_i and its estimation $g(x_i)$ is less than ϵ. This criterion is closely related to the model prediction developed in [22, 23]. There, for a fixed proportion of instances, we aimed to minimize the width of the band models. Here, for a fixed width 2ϵ, we maximize the proportion of instances in the band $(g(\cdot) - \epsilon, g(\cdot) + \epsilon)$.*

Example 4 *Let us consider a binary classification problem. Let $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ denote the 0-1 loss function defined as $\Delta(y, y') = 1_{y \neq y'}$. Under these assumptions, the three stochastic orderings above reviewed (dominance in the sense of expected utility, first stochastic dominance and statistical preference) lead to the same pairwise comparison between classification models. In fact, given a pair of arbitrary models $f : \mathcal{X} \to \mathcal{Y}$ and $f' : \mathcal{X} \to \mathcal{Y}$, we observe that:*

- *$E[U^{(\mathbf{X},Y)}_{\Delta,f}] = -P(Y \neq f(\mathbf{X}))$. In words, the reward expectation coincides with the opposite of the (theoretical) proportion of misclassifications. Therefore, according to the criterion of dominance in the sense of expected utility, f is said to be preferred to f' whenever the proportion of misclassifications of f is less than the respective proportion associated to f'.*

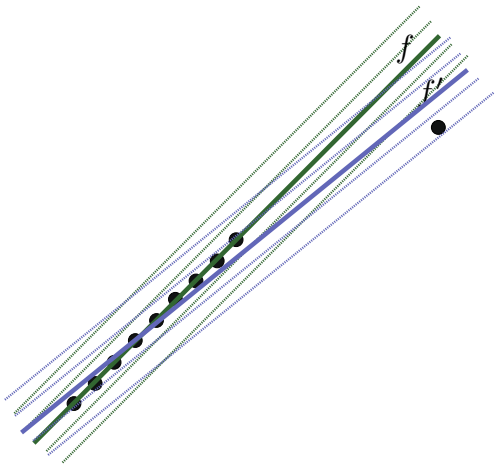

Figure 2: First stochastic dominance wrt square loss. For a small width, model f is better than model f'. For a bigger width, the contrary happens.

- $P(U_{\Delta,f}^{(\mathbf{X},Y)} > c) = \begin{cases} 1 & \text{if } c < -1 \\ P(Y = f(\mathbf{X})) & \text{if } -1 \leq c < 0 \\ 0 & \text{if } c \geq 0, \end{cases}$ *and the same applies to* f'. *Thus, according to the criterion of first stochastic dominance, model* f *is preferred to* f' *whenever it produces a higher proportion of appropriate classifications, or equivalently, a lower proportion of misclassifications.*
- $P(U_{\Delta,f}^{(\mathbf{X},Y)} > U_{\Delta,f'}^{(\mathbf{X},Y)}) = P(Y = f(\mathbf{X}), Y \neq f'(\mathbf{X}))$. *Analogously,* $P(U_{\Delta,f'}^{(\mathbf{X},Y)} > U_{\Delta,f}^{(\mathbf{X},Y)}) = P(Y = f'(\mathbf{X}), Y \neq f(\mathbf{X}))$. *Thus, we can easily deduce that, according to the preference ordering between models derived from statistical preference,* f *is preferred to* f' *whenever the proportion of appropriate classifications is higher with* f *than with* f'.

Example 5 *Let us consider a binary classification problem again, but let us now consider the non-symmetric loss function determined by the following cost matrix:*

$$\begin{pmatrix} 0 & 1 \\ 0.2 & 0 \end{pmatrix}$$

In this situation, the three stochastic orderings above reviewed do not produce the same criterion of comparison between models. Let the two classes be respectively denoted by "1" and "2". Then:

- $E[U_{\Delta,f}^{(\mathbf{X},Y)}] = -P(Y = 1, f(\mathbf{X}) = 2) - 0.2P(Y = 2, f(\mathbf{X}) = 1)$.
- $P(U_{\Delta,f}^{(\mathbf{X},Y)} > c) = \begin{cases} 1 & \text{if } c < -1 \\ 1 - P(Y = 1, f(\mathbf{X}) = 2) & \text{if } -1 \leq c < -0.2 \\ P(Y = f(\mathbf{X})) & \text{if } -0.2 \leq c < 0 \\ 0 & \text{if } c \geq 0, \end{cases}$ *and the same applies to* f'. *Thus, according to the criterion of first stochastic dominance,* f *is preferred to* f' *if* $P(Y = f(\mathbf{X})) \geq P(Y = f'(\mathbf{X}))$ *and furthermore* $P(Y = 1, f(\mathbf{X}) = 2) < P(Y = 1, f'(\mathbf{X}) = 2)$. *This criterion is more restrictive than the above criterion, which is just based on the expected reward.*
- $P(U_{\Delta,f}^{(\mathbf{X},Y)} > U_{\Delta,f'}^{(\mathbf{X},Y)}) = P(Y = f(\mathbf{X}), Y \neq f'(\mathbf{X}))$. *Analogously,* $P(U_{\Delta,f'}^{(\mathbf{X},Y)} > U_{\Delta,f}^{(\mathbf{X},Y)}) = P(Y = f'(\mathbf{X}), Y \neq f(\mathbf{X}))$. *Let the reader notice that this criterion depends on the probability distribution of the 3-dimensional vector* $(Y, f(\mathbf{X}), f'(\mathbf{X}))$, *while the other ones just depend on the respective distributions of the 2-dimensional vectors* $(Y, f(\mathbf{X}))$ *and* $(Y, f'(\mathbf{X}))$.

All the above examples considered a numerical valued loss function Δ. As we have mentioned before, sometimes the expert cannot provide us with such a numerical function, but (s)he can establish a preference ordering on $\mathcal{Y} \times \mathcal{Y}$. Let us slightly modify the Example 5 by replacing the numerical-valued loss function by a non-numerical one.

Example 6 *Consider the binary classification problem from Example 5, but instead of considering a specific numerical loss function, let us suppose that the expert just provides us with the following information:*

$$\Delta(1,1) = \Delta(2,2) < \Delta(1,2) < \Delta(2,1),$$

reflected that misclassifications of elements Class 1 have more negative impact that misclassifications of those instances that actually belong to Class 2. Without defining a particular numerical loss function, dominance in the sense of expected utility cannot be applied. Contrarily, the comparison methods proposed in Definitions 2 and 3 lead exactly to the same criteria illustrated in Example 5.

In practice, we do not need to restrict ourselves to the three above preference pre-orderings between models (dominance based on expected reward, $\preceq_{1st}$ $-\Delta-$ or $\preceq_{sp}$ $-\Delta-$ dominance). According to [1, 2] a general formulation encompasses a wide variety of stochastic orderings from what different orderings between models can be derived:

Definition 4 *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space. A random variable* $Z : \Omega \to \mathbb{R}$ *is preferred to another one* $Z' : \Omega \to \mathbb{R}$ *if and only if*

$$E[F(Z, Z')] \geq E[F(Z', Z)], \tag{4}$$

where $F : \mathbb{R}^2 \to \mathbb{R}$ *is a measurable mapping, increasing in the first component and decreasing in the second one.*

If we consider the mapping $G : \mathbb{R}^2 \to \mathbb{R}$ defined as $G = F - F \circ \text{sw}$ where $\text{sw} : \mathbb{R}^2 \to \mathbb{R}^2$ is the "swap" function $\text{sw}(x, y) = (y, x), \forall (x, y) \in \mathbb{R}^2$, the above generalized preference criterion between measurable mappings can be equivalently formulated as follows: $Z : \Omega \to \mathbb{R}$ is preferred to another one $Z' : \Omega \to \mathbb{R}$ if and only if

$$E[G(Z, Z')] \geq 0. \tag{5}$$

The above definition can be further generalized to the case where the random variables Z and Z' take values on a preposet $\mathcal{O}$ with structure of measurable space (and therefore F, sw and G are defined on $\mathcal{O} \times \mathcal{O}$ instead of $\mathbb{R}^2$). According to these generalized stochastic orderings, Definition 1 can be adapted into two different equivalent formulations:

$$\begin{aligned} &f \text{ is } F - \Delta - \text{preferred to } f' \text{ if} \\ &E[F(U_{\Delta,f}^{(\mathbf{X},Y)}, U_{\Delta,f'}^{(\mathbf{X},Y)})] \geq E[F(U_{\Delta,f'}^{(\mathbf{X},Y)}, U_{\Delta,f}^{(\mathbf{X},Y)})]. \end{aligned} \tag{6}$$

$$\begin{aligned} &f \text{ is } G - \Delta - \text{preferred to } f' \text{ if} \\ &E[G(U_{\Delta,f}^{(\mathbf{X},Y)}, U_{\Delta,f'}^{(\mathbf{X},Y)})] \geq 0. \end{aligned} \tag{7}$$

The usual criterion that compares the expected risks of a pair of models, as well as the comparing criteria proposed in Definitions 2 and 3 are particular instances of Equation 7. In fact, we can recover the three of them by means of the following respective particularizations of F and G:

- *Expected risk criterion:* $F(x, y) = x$ and therefore $G = F - F \circ \text{sw}$ is $G(x, y) = x - y$, $\forall (x, y) \in \mathbb{R}^2$.
- $\preceq_{1st}$ $-\Delta-$ *dominance:* In this case we cannot consider a single function F, but a family of mappings $\{F_c : c \in \mathcal{O}\}$, where $F_c : \mathcal{O} \times \mathcal{O} \to \{0, 1\}$ is defined as $F_c(x, y) = 1_{x >_{\mathcal{O}} c}$, $\forall c$. Therefore $G_c : \mathcal{O} \times \mathcal{O} \to \{-1, 0, 1\}$ is given by the formula $G_c(x, y) = 1_{x >_{\mathcal{O}} c} - 1_{y >_{\mathcal{O}} c}$, for each $c \in \mathcal{O}$. We can easily check that f is $\preceq_{1st}$ $-\Delta-$ preferred to f' if $E[G_c(U_{f,\Delta}, U_{f',\Delta})] \geq 0$, $\forall c \in \mathcal{O}$.
- $\preceq_{sd}$ $-\Delta-$ *dominance:* $F(x, y) = 1_{x >_{\mathcal{O}} y}$ and therefore $G = F - F \circ \text{sw} : \Omega \times \Omega \to \{-1, 0, 1\}$ is given by the expression $G(x, y) = 1_{x >_{\mathcal{O}} y} - 1_{y >_{\mathcal{O}} x}$. In the particular case where numerical rewards are considered, and therefore, $\mathcal{O}$ coincides with the real line, G can be expressed in a more compact way as $G(x, y) = \text{sign}(x - y)$ where "sign" denotes the mapping that takes the values -1, 0 or 1, depending on the sign (negative, null or positive) of the argument.

4. Set-valued data

So far, we have considered the situation where the vector of attributes, $\mathbf{X}$, as well as the response variable, Y, are observed with precision. Now, we will look at the more general case of set-valued datasets. This more general situation has been addressed during the last decades from different perspectives, according to different interpretation of sets. Here we will consider the so called "epistemic" or "disjunctive" interpretation of them ([3, 4, 11]). According to it, sets represent incomplete information about specific elements of the universe ($\mathcal{X} \times \mathcal{Y}$, in our case), containing all (mutually exclusive) possibilities. If the available information about each pair $(\mathbf{X}(\omega), y(\omega))$ is of the form $(\mathbf{X}(\omega), y(\omega)) \in \mathbb{X}(\omega) \times \mathbb{Y}(\omega)$, it means that any value outside the set $\mathbb{X}(\omega) \times \mathbb{Y}(\omega) \subseteq \mathcal{X} \times \mathcal{Y}$ is considered impossible, and only one value in $\mathbb{X}(\omega) \times \mathbb{Y}(\omega)$ is correct. This ill-knowledge about each pair $(\mathbf{X}(\omega), y(\omega))$ can be due to different reasons, some of them described in [12].

In some previous works (see, for instance, [25, 18, 19, 20]), the performance of a model was represented by means of the set of possible values for the risk, i.e., the set:

$$\{E[U_{\Delta,f}^{(\mathbf{X},Y)}] : (\mathbf{X}, Y) \in S(\mathbb{X} \times \mathbb{Y})\},$$

where $S(\mathbb{X} \times \mathbb{Y})$ represents the collection of selections of the multi-valued mapping $\mathbb{X} \times \mathbb{Y} : \Omega \to \wp(\mathcal{X} \times \mathcal{Y})$. The above collection of possible values for the expectation corresponds to the Aumann integral of the multi-valued mapping $U_{\Delta,f}^{\mathbb{X} \times \mathbb{Y}} : \Omega \to \wp(\mathbb{R})$ defined as

$$U_{\Delta,f}^{\mathbb{X} \times \mathbb{Y}}(\omega) = \{-\Delta(y, f(\mathbf{x})) : (\mathbf{x}, y) \in \mathbb{X}(\omega) \times \mathbb{Y}(\omega)\}. \tag{8}$$

Thus, in order to compare the performances of two different models, $f : \mathcal{X} \to \mathcal{Y}$ and $f' : \mathcal{X} \to \mathcal{Y}$ the Aumann expectations of $U_{\Delta,f}^{\mathbb{X} \times \mathbb{Y}}$ and $U_{\Delta,f'}^{\mathbb{X} \times \mathbb{Y}}$ (or their sample estimations) were compared, according to some pre-order between intervals or sets.

Such a comparison between the Aumann expectations of the sets of possible rewards generalize the expected loss minimization criterion, and therefore, it makes use of a generalized version of the criterion of dominance in the sense of expected utility. With this idea in mind, we can easily extend the generalized approach proposed in the last section to the set-valued setting. In order to do so, we will identify every model f with the multi-valued mapping defined in Equation 8. Thus, in order to compare any pair of models, we will combine a stochastic ordering with an ordering between intervals, leading to a generalized stochastic ordering in the sense of [2] and [9].

4.1. Interval (pre)orderings

Let us first enumerate the most common preorderings between closed intervals in the literature. Later we will review some proposals from the literature where those pre-orderings between intervals have been used to compare ill-known expected rewards, and in turn, to select the best model (or a

set of non-dominated ones, instead). We will divide the collection of orderings between intervals into two groups: those that produce a total order, and those others admitting that some intervals my be incomparable due to imprecision. The first family includes:

- Weak ordering: $[a_1, b_1] \leq_{wo} [a_2, b_2]$ if $a_1 \leq b_2$
- Maximin [29]: $[a_1, b_1] \leq_{Mm} [a_2, b_2]$ if $a_1 \leq a_2$
- Maximax [27]: $[a_1, b_1] \leq_{MM} [a_2, b_2]$ if $b_1 \leq b_2$
- Hurwicz [16]: $[a_1, b_1] \leq_{H(\gamma)} [a_2, b_2]$ if $\gamma a_1 + (1-\gamma)b_1 \leq \gamma a_2 + (1-\gamma)b_2$ where $\gamma \in [0, 1]$.

Examples of the second family (partial preorders) are the following ones:

- Interval dominance [13]: $[a_1, b_1] \leq_{id} [a_2, b_2]$ if $a_1 \leq a_2$.
- lattice ordering: $[a_1, b_1] \leq_{lo} [a_2, b_2]$ if $a_1 \leq a_2$ and $b_1 \leq b_2$.

Interval dominance is the most restrictive criterion, and according to it two intervals are only comparable when they do not overlap, while the lattice ordering is the canonical order induced by the lattice of intervals (i.e. $[a_1, b_1] \leq_{lo} [a_2, b_2]$ iff $\max([a_1, b_1], [a_2, b_2]) = [a_2, b_2]$, or equivalently iff $\inf([a_1, b_1], [a_2, b_2]) = [a_1, b_1]$).

4.2. Model selection based on interval (pre)orderings

The above pre-orders are usually formulated for closed intervals, but they can be also used to compare any pair of bounded subsets of the real line (non necessarily convex or closed), if we consider their respective convex closed hulls. In [25], the criterion of interval dominance has been used, and therefore a model f was said to be preferred to another one f' whenever the following inequality holds: the expectation of the supremum of the multivalued mapping $U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f'}$ is less than the infimum of the expectation of $U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f}$. Thus, if the Aumann expectation of $U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f}$ dominates the respective expectation of $U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f'}$ then we can assure that, whatever the selection $(\mathbf{X}, y)$ of $\mathbb{X} \times \mathbb{Y}$ is, the expected reward of f is greater than expected reward of f'. Also, a pair of independent uniform distributions over the respective sets of possible values for the expectation rewards are considered, in order to derive a total ordering between models, and Hurwicz criterion with $\gamma = 0.5$ has been considered. In [15], the maximin criterion is used to select the (set of) optimal model(s) in a parametric family. This interval ordering criterion naturally arises on the basis of some additional assumptions about the model. In [28], authors provide explicit formulas for the empirical risk under the minimin and minimax strategies (the maximin and maximax strategies, under our dual setting) for different loss functions in some regression and classification problems.

In our generalized context, the images of set-valued reward functions $U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f}$ and $U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f'}$ are not necessarily subsets of the real line, but subsets of a general preposet, $\mathcal{O}$. Thus, alternatives to the comparison of their respective Aumann expectations must be considered. In last section, we proposed such a generalized procedure, but just for the case where the samples were observed with precision. There we replaced the traditional inequality between reward expectations, i.e., the inequality:

$$E\left[U^{(\mathbf{X},Y)}_{\Delta,f}\right] \geq E\left[U^{(\mathbf{X},Y)}_{\Delta,f'}\right] \tag{9}$$

or the equivalent inequality:

$$E\left[U^{(\mathbf{X},Y)}_{\Delta,f} - U^{(\mathbf{X},Y)}_{\Delta,f'}\right] \geq 0 \tag{10}$$

by the more general comparisons proposed in Equations 6 and 7 (respectively generalizing Equations 9 and 10). According to the (equivalent) inequalities proposed in Equations 6 and 7, we could identify any possible (pre)ordering defined over the family of models with a pair (Δ, F), constituted by a specific loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathcal{O}$ and a specific stochastic ordering determined by a mapping $F : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ (or the associated mapping $G : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ defined as $G = F - F \circ \mathrm{sw}$). We have illustrated some of those (pre)orderings between models in Examples 2 to 6.

Now, with interval data, the respective rewards of f and f', $U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f}$ and $U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f'}$, are multi-valued mappings, and therefore, the same happens with their images through F or G.

Natural generalizations of Equations 6 and 7 to the case of set-valued data are as follows:

$$\begin{aligned} &f \text{ is } F-\Delta-\text{ord}-\text{preferred to } f' \text{ if} \\ &\quad E[F(U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f})] \geq_{\text{ord}} E[F(U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f'})]. \end{aligned} \tag{11}$$

$$\begin{aligned} &f \text{ is } G-\Delta-\text{preferred to } f' \text{ if} \\ &\quad E[G(U^{\mathbb{X}\times\mathbb{Y}}_{\Delta,f})] \geq_{\text{ord}} 0, \end{aligned} \tag{12}$$

where "ord" represents one of the (pre)orders between bounded sets listed in Subsection 4.1.

5. Concluding remarks

We have established a common formal framework that includes different comparison methods between pairs of regression or classification modes from set-valued data. In this context, for each triplet (loss function, stochastic ordering, interval ordering), we get a particular model comparison criterion. Several particular instances of the general framework introduced here have been independently proposed by different authors in recent papers. Up to our knowledge, all of them are based on stochastic dominance in the sense of expected utility. In our general framework other comparisons between models, including the case where the expert does not provide numerical utility/loss functions, also make sense.

Acknowledgments

This work has been partially supported by the Spanish projects TIN2011-24302 and TIN2014-56967-R, and also by the Labex CIMI at Paul Sabatier University.

References

[1] I. Couso, Preference Relations and Families of Probabilities: Different Sides of the Same Coin, IPMU 2012, Communications in Computer and Information Science, Vol. 442, 2014, pp 1-9.

[2] I. Couso, D. Dubois, An imprecise probability approach to joint extensions of stochastic and interval orderings, Advances in Computational Intelligence, pp. 388-399. Springer, 2012.

[3] I. Couso, D. Dubois, Statistical reasoning with set-valued information: Ontic vs. epistemic views. International Journal of Approximate Reasoning 55 (2014)1502-1518.

[4] I. Couso, D. Dubois, L. Sánchez, Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables, Springer, 2014.

[5] I. Couso, S. Moral, L. Sánchez, The behavioral meaning of the median (2015) Information Sciences 294, 127–138.

[6] H. David, The method of paired comparisons, Griffin's Statistical Monographs & Courses, vol. 12, Charles Griffin & D. Ltd., London, 1963.

[7] T. Denœux, Maximum likelihood estimation from fuzzy data using the EM algorithm, Fuzzy Sets and Systems 183 (2011) 72–91.

[8] T. Denœux, Maximum likelihood estimation from uncertain data in the belief function framework, IEEE Trans. Knowl. Data Eng. 25 (2013) 119–130.

[9] S. Destercke, I. Couso, Ranking of fuzzy intervals seen through the imprecise probabilistic lens, Fuzzy Sets and Systems, d.o.i. 10.10106/j.fss.2014.12.009.

[10] D. Dubois, Ontic vs. epistemic fuzzy sets in modeling and data processing tasks, in: K. Madani, J. Kacprzyk, J. Filipe (Eds.), Proc. IJCCI (NCTA), International Conference on Neural Computation Theory and Applications, Paris, 2011.

[11] D. Dubois, H. Prade, Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets, Fuzzy Sets and Systems 192 (2012) 3–24.

[12] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf and L. Ginzburg, Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty, Sandia National Laboratories, Report SAND2007-0939 (2007).

[13] Fishburn, P.C., Interval Orderings and interval graphs, Wiley, New-York, 1985.

[14] J. Hadar, W. Russell, Rules for Ordering Uncertain Prospects, American Economic Review 59, March 1969, 25-34.

[15] E. Hüllermeier, Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization, International Journal of Approximate Reasoning 55 (2014) 1519–1534.

[16] L. Hurwicz, A class of criteria for decision-making under ignorance, Cowles Comission Paper, 356, 1951.

[17] A.M. Palacios, L. Sánchez, I. Couso, Diagnosis of dyslexia with low quality data with genetic fuzzy systems, International Journal of Approximate Reasoning Volume 51 (2010) 993-1009.

[18] A.M. Palacios, L. Sánchez, I. Couso, Linguistic cost-sensitive learning of Genetic Fuzzy Classifiers for imprecise data, International Journal of Approximate Reasoning Volume 52 (2011) 841–862.

[19] A.M. Palacios, L. Sánchez, I. Couso, Future Performance Modeling in Athletism with Low Quality Data-based Genetic Fuzzy Systems, Journal of Multiple-Valued Logic and Soft Computing 17 (2011) 207–228.

[20] A.M. Palacios, L. Sánchez, I. Couso, Boosting of Fuzzy Rules with Low Quality Data, Journal of Multiple-Valued Logic and Soft Computing 19 (2012) 591–619.

[21] L.J. Savage, The Foundations of Statistics. Wiley (1954); 2nd edition, Dover Publications Inc., New York, 1972.

[22] L. Sánchez, Interval-valued GA-P algorithms, IEEE Transactions on Evolutionary Computation 4 (2000) 64–72.

[23] L. Sánchez, I. Couso, Learning from imprecise examples with GA-P algorithms, Mathware and Soft Computing 5 (1998) 305-319.

[24] L. Sánchez, I. Couso, Advocating the use of imprecisely observed data in genetic fuzzy systems, IEEE Trans. Fuzzy Syst. 15 (4) (2007) 551-562.

[25] L. Sánchez, I. Couso, J. Casillas, Genetic learning of fuzzy rules based on low quality data, Fuzzy Sets and Systems 160 (2009) 2524–2552.

[26] L. Sánchez, J. Otero, I. Couso, Obtaining linguistic fuzzy rule-based regression models from imprecise data with multiobjective genetic algorithms, Soft Computing 13 (2009) 467–479.

[27] L. E. Satia, J. Roy, Markovian decision processes with uncertain transition probabilities. Operations Research, 21 (1973) 728–740.

[28] L. Utkin, F. P.A. Coolen, Interval-valued regression and classification models in the framework of machine learning, 7th International Symposium on Imprecise Probability: Theories and Applications, Innsbruck, Austria, 2011.

[29] A.Wald, Statistical decision functions which minimize the maximum risk, Annals of Mathematics 46 (1945) 265–280.