Practical notes on applying generalised stochastic orderings to the study of performance of classification algorithms for low quality data

Patryk Żywica[1,⋆], Katarzyna Basiukajc[1], and Inés Couso[2]

[1] Faculty of Mathematics and Computer Science
Adam Mickiewicz University in Poznań, Poland
Umultowska 87, 61-614 Poznań, Poland
bikol@amu.edu.pl

[2] Department of Statistics and Operational Research
University of Oviedo, Spain
C/ Luis Berrocal s/n, Despacho 4.1.12, 33071 Gijón, Spain
couso@uniovi.es

Abstract. This paper presents an approach to applying stochastic orderings to evaluate classification algorithms for low quality data. It discusses some known stochastic orderings along with practical notes about their application to classifier evaluation. Finally, a new approach based on fuzzy cost function is presented. The new method allows comparing any two classifiers, but does not require a precise definition of the cost function. All proposed methods were evaluated on real life medical data. The obtained results are very similar to those previously reported but comparatively much weaker assumptions about costs values are adopted.

Keywords: Classification, Loss function, Stochastic ordering, Low quality data, Fuzzy random variable

Acknowledgement. This work was partially supported by the Polish National Science Centre grant number 2016/21/N/ST6/00316.

1 Introduction

As long as machine learning algorithms are becoming more and more popular and their area of application is simultaneously expanding, we are facing a wide range of newly arising problems. One of them is the evaluation of algorithms concerning real-life conditions with some unusual restrictions.

A typical binary classification problem's goal is to find a model $f : \mathcal{X} \to \mathcal{Y}$ assigning the categories from $\mathcal{Y} = \{0, 1\}$ to lists of attributes mapped by $\mathbb{Y} : \mathit{\Omega} \to \mathcal{Y}$ and $\mathbb{X} : \mathit{\Omega} \to \mathcal{X}$, respectively. These kinds of models can be evaluated by widely known evaluation functions such as: accuracy, precision, recall as well as F1-score.

[⋆] Corresponding author

Unfortunately, not every classification problem matches the definition above [1,2,3,4]. For instance, in some medical diagnostic problems, the physician is not always able to collect all the data needed for the diagnosis due to the time and money investment this constitutes [5,6]. In such situations we may want to evaluate classifiers similarly to the way real-life doctor's decisions are evaluated. We thus, take into account the data quality, the level of uncertainty, and allow the possibility of receiving a "not available" (NA) value as the output of the classification model. In turn, this prevent the use of mentioned evaluation functions so another solution must be found.

One of the proposed approaches is to introduce the cost function matching model outcomes (projected as *true positive*, *true negative*, etc.) with the cost of the real-life consequences they can contribute to [7]. Although this solution works, the selection of cost matrix values is subjective with possibly divergent opinions held by experts. Furthermore, some small changes in the cost matrix values may cause significant changes in the final classification result (lack of robustness). Another problem related especially to medical decision evaluation is that costs of individual decisions may not be known. For example, for each patient the cost of *false–negative* may be different depending on his or her other medical conditions [8].

Another idea to study the performance of such classification algorithms is to apply stochastic orderings. This method was proposed in [9,10] and it fits into the situation presented. It will be presented and extended in following sections.

The remainder of the paper is organised as follows. In Section 2 we present basic notions regarding cost function and stochastic orders. Section 3 describes an evaluated dataset as well as results for three stochastic order based classifier evaluation methods. In section 4 we present details of our proposed approach as well as some analysis of obtained results. Conclusions and further work appear in Section 5.

2 Basic notions

Let $\mathcal{Y}$ be the output space of the models and let $\varDelta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be the *loss (cost) function* used to find the best classifier. The main goal of the loss function is to penalise wrong outcomes in order to enable finding the best model as the model with the minimal loss. It means, that loss function values usually become positive when the predictions do not match reality.

There also exists a dual definition of the *loss function* called *reward function* which we will use in this paper to match the common stochastic orderings literature. It's a simple opposite to the loss function which means the greater value of the reward function is, the more relevant outcomes are. Therefore, the best model can be found at its maximum. We can associate the random variable *reward* $U^{(\mathbb{X},\mathbb{Y})}_{\Delta,f} : \varOmega \to \mathbb{R}$ with the model f using the definition as follows:

$$U^{(\mathbb{X},\mathbb{Y})}_{\Delta,f} = -\varDelta(\mathbb{Y}(\omega), f(\mathbb{X}(\omega)) \qquad \forall \omega \in \varOmega. \tag{1}$$

Table 1: Cost function (matrix) for example binary classification (a) and its extension to uncertain classification (b).

		predicted	
		benign	malignant
actual	benign	TN: 0	FP: 2.5
	malignant	FN: 5	TP: 0

(a)

	predicted		
	benign	malignant	NA
benign	TN: 0	FP: 2.5	N0: 1
malignant	FN: 5	TP: 0	N1: 2

(b)

Let $X, Y : \Omega \to \mathbb{R}$ be two random variables defined on the same probability space $(\Omega, \mathcal{A}, P)$. From many available kinds of stochastic orderings, of most interest from classification performance evaluation point of view are [11]:

1. Dominance in the sense of expected utility [12]. Given an increasing function $u : \mathbb{R} \to \mathbb{R}$, X dominates Y wrt u (denoted $X \succeq_u Y$) if
$$E_P(u(X)) \geq E_P(u(Y))\,. \tag{2}$$
2. First order stochastic dominance [13]. X dominates Y (denoted $X \succeq_{1st} Y$) if
$$\forall_{x \in \mathbb{R}} \quad P(X > x) \geq P(Y > x)\,. \tag{3}$$
It is well known that $X \preceq_{1st} Y$ if and only if $X \preceq_u Y$, for all increasing utility functions $u : \mathbb{R} \to \mathbb{R}$.
3. Statistical preference [14]. X is statistically preferred to Y (denoted $X \succeq_{sp} Y$) if
$$P(X > Y) \geq P(Y > X)\,. \tag{4}$$

Based on this we can present the notion of the $(\succeq, \Delta)$*–domination* proposed by Couso and Sánchez [9].

Definition 1. *Let $f_1 : \mathcal{X} \to \mathcal{Y}$ and $f_2 : \mathcal{X} \to \mathcal{Y}$ be the classification models and $\succeq$ be any stochastic ordering. f_1 $(\succeq, \Delta)$–dominates f_2 if*
$$U^{(\mathbb{X},\mathbb{Y})}_{\Delta,f_1} \succeq U^{(\mathbb{X},\mathbb{Y})}_{\Delta,f_2}\,. \tag{5}$$

Example 1 *Let us consider the binary classification problem which refers to determination whether the tumor is malignant (M) or benign (B). Let $\mathbb{X}$ be the random vector of attributes and $\mathbb{Y}$ – the outcome. Let the cost matrix be given as in Table 1a.*

According to the definition of the reward function we have:
$$P(U^{(\mathbb{X},\mathbb{Y})}_{\Delta,f} > c) = \begin{cases} 1, & \text{if } c < -5, \\ 1 - P(\mathbb{Y} = M, f(\mathbb{X}) = B), & \text{if } -5 \leq c < -2.5, \\ P(\mathbb{Y} = f(\mathbb{X})), & \text{if } -2.5 \leq c < 0, \\ 0, & \text{if } c \geq 0. \end{cases} \tag{6}$$

The equation above together with Definition 1 lead to the conclusion that f_1 $(\succeq_{1st}, \varDelta)$–dominates f_2 if and only if:

$$P(\mathbb{Y} = M, f_1(\mathbb{X}) = B) \leq P(\mathbb{Y} = M, f_2(\mathbb{X}) = B), \tag{7}$$

$$P(\mathbb{Y} = f_1(\mathbb{X})) \geq P(\mathbb{Y} = f_2(\mathbb{X})). \tag{8}$$

Moreover, this stays true for any cost matrix for which

$$\varDelta(\mathbb{Y} = M, f(\mathbb{X}) = B) \geq \varDelta(\mathbb{Y} = B, f(\mathbb{X}) = M) \geq \varDelta(\mathbb{Y} = f(\mathbb{X}))\,. \tag{9}$$

This means that $(\succeq_{1st}, \varDelta)$–dominance does not depend on actual cost values but only on their order.

Example 2 *Let us consider again the binary classification problem from Example 1 with the same cost matrix. We can distinguish following cases:*

$$P(U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_1} > U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_2}) = P(f_1(\mathbb{X}) = B, f_2(\mathbb{X}) = M), \quad if \quad \mathbb{Y} = B, \tag{10}$$

$$P(U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_1} > U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_2}) = P(f_1(\mathbb{X}) = M, f_2(\mathbb{X}) = B), \quad if \quad \mathbb{Y} = M. \tag{11}$$

According to the (10) and (11) and Definition 1, f_1 $(\succeq_{sp}, \varDelta)$–dominates f_2 if and only if:

$$P(\mathbb{Y} = f_1(\mathbb{X}), \mathbb{Y} \neq f_2(\mathbb{X})) \geq P(\mathbb{Y} = f_2(\mathbb{X}), \mathbb{Y} \neq f_1(\mathbb{X})). \tag{12}$$

Example 3 *Let consider medical classification problem for* 10 *patients, cost matrix from Example 1 and three diagnostic models f_1, f_2 and f_3. Actual diagnoses and predictions are given in Table 2. We can easily calculate that*

$$P(\mathbb{Y} = f_1(\mathbb{X})) = 0.8 \qquad P(\mathbb{Y} = M, f_1(\mathbb{X}) = B) = 0.1 \tag{13}$$

$$P(\mathbb{Y} = f_2(\mathbb{X})) = 0.6 \qquad P(\mathbb{Y} = M, f_2(\mathbb{X}) = B) = 0 \tag{14}$$

$$P(\mathbb{Y} = f_3(\mathbb{X})) = 0.7 \qquad P(\mathbb{Y} = M, f_3(\mathbb{X}) = B) = 0.2 \tag{15}$$

According to Example 1, f_1 $(\succeq_{1st}, \varDelta)$–dominates f_3. Unfortunately, models f_1 and f_2 are incomparable with the respect to $(\succeq_{1st}, \varDelta)$ criterion.

Moreover, according to Example 2,

$$P(U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_1} > U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_2}) = 0.3 \qquad P(U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_2} > U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_1}) = 0.1 \tag{16}$$

$$P(U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_1} > U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_3}) = 0.1 \qquad P(U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_3} > U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_1}) = 0 \tag{17}$$

thus f_1 $(\succeq_{sp}, \varDelta)$–dominates f_2 as well as f_1 $(\succeq_{sp}, \varDelta)$–dominates f_3.

Table 2: Diagnoses for patients from Example 3.

	Patients									
Diagnosis	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
actual $\mathbb{Y}(\omega)$	B	B	B	B	B	M	M	M	M	M
model $f_1(\mathbb{X}(\omega))$	B	B	B	B	M	M	M	M	M	B
model $f_2(\mathbb{X}(\omega))$	B	M	M	M	M	M	M	M	M	M
model $f_3(\mathbb{X}(\omega))$	B	B	B	B	M	M	M	M	B	B

3 Application of stochastic orderings to low quality data classification performance evaluation

Definitions and examples presented in Section 2 referred to binary classification problem. However, as it was mentioned in the Introduction, there exist real-life problems, where the classification can be uncertain and the outcome may take the *NA* value. In this section we are going to apply concepts from Section 2 to this particular case.

3.1 Medical data

We base our evaluation on test dataset from recent research on application of aggregation operators to incomplete data classification [2]. Original study group consists of 388 patients diagnosed and treated for ovarian tumor in the Division of Gynecological Surgery, Poznan University of Medical Sciences, between 2005 and 2015. Among them, 61% were diagnosed with a benign tumor and 39% with a malignant one. Moreover, 56% of the patients had no missing values in the attributes required by diagnostic models, 40% had a percentage of missing values in the range $(0\%, 50\%]$, and the remainder had more than 50% missing values. The test set consists of patients with real missing data and some proportion of patients with a complete set of features. As a result, the test set consisted of 175 patients. Patients with more than 50% missing values were excluded from the study. The dataset partition is presented visually in Figure 1.

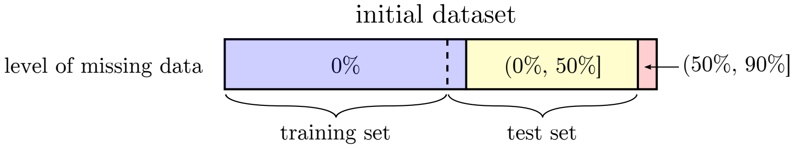


Fig. 1: The division of the dataset. Patients with more than 50% missing values were not included in the experiment. Source [2].

During the research over 4000 different classification strategies were evaluated. Among them 130 were selected into test phase. Our evaluation was performed on outcomes returned by those classifiers on real life test set. For all classifiers it was assumed that no diagnosis may be returned. For more information regarding dataset we refer the reader to original paper [2].

3.2 Expected utility

One of the methods of dealing with evaluation of the algorithms based on incomplete data is to insert an additional column to the cost matrix as it was proposed in [2]. Then, the loss function may be defined as a sum of costs of all outcomes given by the algorithm. Unfortunately, it's hard to assume that one type of mistake is a certain number of times worse than the other. We also can't say, that every patient has the same loss for every mistake. These make the costs only intuitive and as long as their small changes lead to different final results, we can't be sure that these final results are the best solutions for the patients.

Table 6 presents some selected best classifiers based on this criterion from the dataset described in previous subsection. As can be seen this comparison method is very useful and straightforward. It offers easy to interpret linear order that facilitates selection of the best classifier. The main drawback of this approach concerns the uncertain and subjective selection of cost function. It it possible that small change to cost matrix causes significant changes to classifier order and this kind of behaviour is not desired.

3.3 First stochastic dominance

To define first of discussed relations, let f be the classifier and Δ the loss function which can be described by the cost matrix similar to one from Table 1b. As the precise cost values do not matter as long as the ordering is saved, let's only assume that the Δ always fulfil (18-21).

$$\Delta(\mathbb{Y} = M, f(\mathbb{X}) = B) \geq \Delta(\mathbb{Y} = B, f(\mathbb{X}) = M) \tag{18}$$
$$\Delta(\mathbb{Y} = B, f(\mathbb{X}) = M) \geq \Delta(\mathbb{Y} = M, f(\mathbb{X}) = NA) \tag{19}$$
$$\Delta(\mathbb{Y} = M, f(\mathbb{X}) = NA) \geq \Delta(\mathbb{Y} = B, f(\mathbb{X}) = NA) \tag{20}$$
$$\Delta(\mathbb{Y} = B, f(\mathbb{X}) = NA) \geq \Delta(\mathbb{Y} = f(\mathbb{X})) \tag{21}$$

Then, analogously to the Example 1, we can conclude, that classifier f_1 $(\succeq_{1st}, \Delta)$–dominates f_2 if and only if:

$$P_{f_1}(TP) + P_{f_1}(TN) \geq P_{f_2}(TP) + P_{f_2}(TN), \tag{22}$$
$$P_{f_1}(TP) + P_{f_1}(TN) + P_{f_1}(N0) \geq P_{f_2}(TP) + P_{f_2}(TN) + P_{f_2}(N0), \tag{23}$$
$$1 - P_{f_1}(FN) - P_{f_1}(FP) \geq 1 - P_{f_2}(FN) - P_{f_2}(FP), \tag{24}$$
$$1 - P_{f_1}(FN) \geq 1 - P_{f_2}(FN). \tag{25}$$

where TP, TN, $N0$, FP, FN are clarified in Table 1b.

Let's take the dataset described in Section 3.1 with $(\succeq_{1st}, \Delta)$–dominance relation. Based on the conditions (22-25), we can define a stochastic ordering inside this set. The only change, applied to make this relation irreflexive as well

Table 3: Diagnoses for patients from Example 4.

	Patients				
Diagnosis	ω_1	ω_2	ω_3	ω_4	ω_5
actual	M	M	M	M	M
model f_1	M	M	*NA*	B	B
model f_2	B	B	M	*NA*	*NA*
model f_3	*NA*	*NA*	B	M	M

as to avoid cycles, is that the condition *greater than* instead of *greater or equal to* must be fulfilled in at least one of (22-25).

Then, we get the strict partial order which allows us to find maximal elements in the set and as we know they are always better than dominated ones, they can be use as an output to further considerations.

Figure 2a and Table 6 shows the maximal elements from the medical classifiers set along with their costs calculated according to the cost matrix from Table 1b and the information if they are the only ones in the chains they are included or not.

Unfortunately, the number of maximal elements is about one-quarter of all models (33 of 130) so this method couldn't help in determining the best classifier but still it can be used as a very effective process of pre-selection.

3.4 Statistical preference stochastic dominance

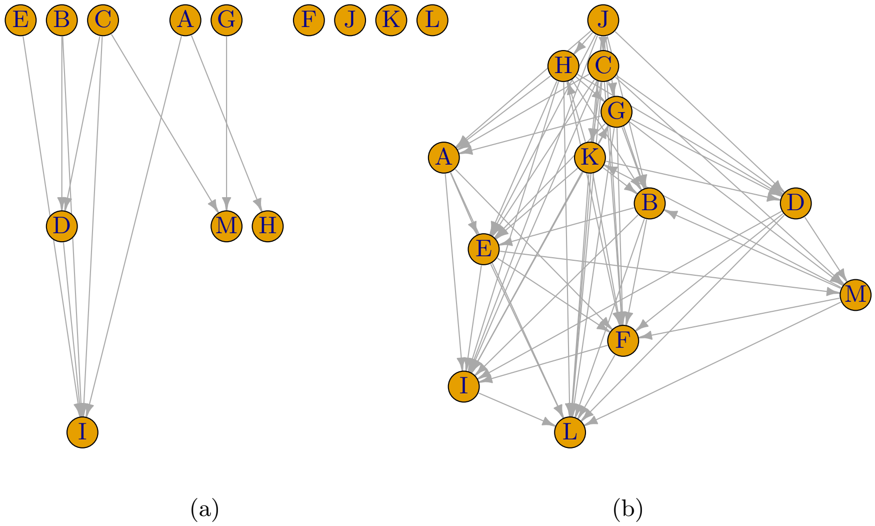
In traditional binary classification, Definition 1 leads to the conclusion that $(\succeq_{sp}, \varDelta)$–domination depends only on the bigger number of true outcomes given by one of the classifiers. In uncertain classification with possibility of NA the problem becomes more complicated. For example, for $\mathbb{Y} = M$ there are three cases when $P(U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_1} > U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_2})$:

- $f_1(\mathbb{X}) = M$ and $f_2(\mathbb{X}) = NA$,
- $f_1(\mathbb{X}) = M$ and $f_2(\mathbb{X}) = B$,
- $f_1(\mathbb{X}) = NA$ and $f_2(\mathbb{X}) = B$.

Similarly, for $\mathbb{Y} = B$, $P(U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_1} > U^{(\mathbb{X},\mathbb{Y})}_{\varDelta,f_2})$ when:

- $f_1(\mathbb{X}) = B$ and $f_2(\mathbb{X}) = NA$,
- $f_1(\mathbb{X}) = B$ and $f_2(\mathbb{X}) = M$,
- $f_1(\mathbb{X}) = NA$ and $f_2(\mathbb{X}) = M$.

Summarising the cases above, we can say that in uncertain classification f_1 $(\succeq_{sp}, \varDelta)$–dominates f_2 if the number of times when f_1 gives proper output while f_2 doesn't or f_1 gives NA while f_2 is wrong is bigger than the number of opposite situations.



(a) (b)

Fig. 2: Graphs showing domination relation for First Stochastic Dominance (a) and Statistical Preference (b) for selected classifiers.

Example 4 *Let's consider medical classification problem for five patients and three diagnostic models f_1, f_2 and f_3 with actual diagnoses and predictions given in Table 3. We can easily notice, that f_2 has three better predictions than f_1, f_3 has three better predictions than f_2 and finally f_1 has three also better predictions than f_3. It means, that f_2 $(\succeq_{sp}, \varDelta)$–dominates f_1, f_3 $(\succeq_{sp}, \varDelta)$–dominates f_2 as well as f_1 $(\succeq_{sp}, \varDelta)$–dominates f_3.*

The previous example shows that in this case statistical preference based dominance relation makes cycles possible to appear while comparing models. It means, that we can compare each pair of classifiers, but it can be impossible to find the maximal elements in the set with $(\succeq_{sp}, \varDelta)$–dominance relation, so the another approach to evaluate the models using this dominance should be defined.

To evaluate models in this case we propose method based on PageRank algorithm [15]. This algorithm is generally used to rate values of the websites by looking how many other sites have reference links to them and how high are the rates of these linking sites. The more links from sites with high rate, the better. All computations are performed on matrix representing graph, where sites are vertices and links are the directed edges pointing to the linked sites. In our situation, models are treated as vertices, the directed edge points to the dominating model and all computations are preserved.

The final score presented in Table 6 is the percentage of time spent in particular classifier vertex while making random PageRank walk. Figure 2b shows the original statistical preference graph for selected best classifiers along with their costs calculated according to the cost matrix from Table 1b.

4 Proposed approach

4.1 Idea

As can be seen from previous section, each presented method has it own strengths and weaknesses. On the one hand, total cost method gives linear order between all classifiers at the expense of the need to provide concrete numerical cost values. On the other hand, First Stochastic Dominance requires only to know whether one classification outcome is better than other. But this leads to a situation where there are many models that cannot be compared. Application of Statistical Preference results with hard to interpret structure. Although application of Page Rank algorithm gives linear order it is still hard to justify such approach and interpret particular values.

Our aim is to propose a method that retains the ability to compare nearly all classifiers, while imposing the least restrictions on the cost of particular, possibly uncertain, decision. As a starting point we chose the First Stochastic Dominance comparison method, which is highly intuitive and easy to interpret. It can be viewed as a total cost method applied for all possible cost functions [16]. Since, experts are often unable to give precise numerical costs, we propose to model them as fuzzy numbers interpreted, in epistemic way (see [17]), as family of nested confidence sets

$$\widetilde{\Delta} : \mathcal{Y} \times \mathcal{Y} \to \mathcal{FN}(\mathbb{R})\,. \tag{26}$$

As will be shown further in this Section, this will enable comparison of all classifiers with respect to any stochastic dominance.

This approach has one additional benefit. Previously cost values were independent of particular patient and were based only on actual and predicted diagnosis. In real life medical scenario this is not always true. For some patients even proper diagnosis may lead to bad outcome and vice versa. Thanks to this approach actual cost corresponding to diagnosis may vary depending on particular patient conditions as we interpret fuzzy number in epistemic way.

4.2 Definitions

Similarly as in previous sections, for any classification model f we can define *reward fuzzy random variable*

$$\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f} : \Omega \to \mathcal{FN}(\mathbb{R}) \tag{27}$$

as a opposite of cost value:

$$\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f} = -\widetilde{\Delta}(\mathbb{Y}(\omega), f(\mathbb{X}(\omega))) \qquad \forall \omega \in \Omega\,. \tag{28}$$

According to the epistemic interpretation, the reward fuzzy random variable should be also understood in terms of confidence sets.

We will use the Extension Principle based stochastic order proposed by Couso and Dubois [16] for comparing fuzzy sets of random variables. Let $\pi_{\widetilde{X}}(X)$ be the

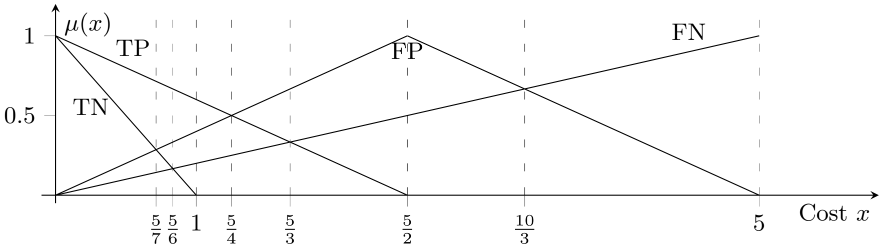


Fig. 3: Fuzzy cost values from Example 5.

degree of possibility that X is the random variable underlain by the fuzzy random variable $\widetilde{X}$

$$\pi_{\widetilde{X}}(X) = \inf_{\omega \in \Omega} \mu_{\widetilde{X}(\omega)}(X(\omega))\,. \tag{29}$$

Then for any stochastic order $\succeq$ the degree of possibility of dominance between fuzzy random variables $\widetilde{X}$ and $\widetilde{Y}$ can be defined as:

$$\Pi(\widetilde{X} \succeq \widetilde{Y}) = \sup_{X,Y\colon X \succeq Y} \min(\pi_{\widetilde{X}}(X), \pi_{\widetilde{Y}}(Y))\,. \tag{30}$$

Now we are ready to define our proposed approach to classification model comparison.

Definition 2. *Let $f_1 : \mathcal{X} \to \mathcal{Y}$ and $f_2 : \mathcal{X} \to \mathcal{Y}$ be the classification models. The degree in which f_1 dominates f_2 with respect to stochastic order $\succeq$ and fuzzy cost function $\widetilde{\Delta}$ is defined as*

$$[\![f_1 \succeq f_2]\!]_{\widetilde{\Delta}} = \frac{\Pi(\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_1} \succeq \widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_2})}{\Pi(\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_1} \succeq \widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_2}) + \Pi(\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_2} \succeq \widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_1})}\,. \tag{31}$$

Such definition, in contrast to simple $[\![f_1 \succeq f_2]\!] = \Pi(\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_1} \succeq \widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_2})$, ensures some desired properties such as $[\![f_1 \succeq f_2]\!] + [\![f_2 \succeq f_1]\!] = 1$ or $[\![f \succeq f]\!] = 0.5$. Moreover, normalisation allows to limit the impact of incomparable random variables when stochastic ordering is a partial preorder. If there are more than two classifiers, we can order them according to maximal degree of being dominated by any other classifier defined for each f_i:

$$p_{\widetilde{\Delta},\succeq}(f_i) = \max_{\substack{\text{all classifiers } f \\ f \neq f_i}} [\![f \succeq f_i]\!]_{\widetilde{\Delta}} \tag{32}$$

When applied stochastic ordering $\succeq$ is a total preorder, then this criterion coincides with the selection of the model f that minimises $\Pi(\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f} \succeq \widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_i})$ – the possibility of being dominated by some (arbitrary) model f_i. Thus (31–32) can be seen as a generalisation of that criterion to partial preorders for which not necessarily $\max(\Pi(\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_1} \succeq \widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_2}), \Pi(\widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_2} \succeq \widetilde{U}^{(\mathbb{X},\mathbb{Y})}_{\widetilde{\Delta},f_1})) = 1$.

Table 4: Degree of domination for classification models from Example 5.

	f_1	f_2	f_3
f_1	0.5	0.7	0.75
f_2	0.3	0.5	0.34
f_3	0.25	0.66	0.5

Table 5: Costs that maximise degree of domination $[\![f_1 \succeq f_2]\!]_{\widetilde{\Delta}}$.

	Patients									
Cost	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
actual $\mathbb{Y}(\omega)$	B	B	B	B	B	M	M	M	M	M
model f_1	0 (B)	0 (B)	0 (B)	0 (B)	2.5 (M)	0 (M)	0 (M)	0 (M)	0 (M)	3.33 (B)
model f_2	0 (B)	3.33 (M)	2.5 (M)	2.5 (M)	2.5 (M)	0 (M)	0 (M)	0 (M)	0 (M)	0 (M)

Example 5 *Let's try to examine the situation from Example 3 using the proposed approach. In the example we will use fuzzy cost function $\widetilde{\Delta}$ with costs defined on Figure 3. The kernels of fuzzy cost values are the same as costs from Example 1.*

Domination degrees are presented in Table 4. Using criterion from (32) we can obtain the following order: f_1 (0.3), f_2 (0.7) and f_3 (0.75). Hence, f_1 is definitely the best model for given problem.

Let us now look in more detail at the situation of f_1 and f_2 models. They were incomparable according to classical First Stochastic Dominance order. Thanks to proposed approach, we still are able to find out which one is better. According to (30) we need to find random variables X and Y that maximise given formula and for which $X \succeq_{1st} Y$ holds. Optimal random variables are given in Table 5. It is easy to observe that for patients for which f_1 outcome was worse then that of f_2 costs are swapped to keep the $X \succeq_{1st} Y$ property.

4.3 Evaluation

We evaluated this approach on the same medical data set as the original classifier comparison strategies. The procedure was following:

1. Extend fuzzy cost function from Example 5 to cover "NA" cases
2. For each pair of classifiers (f_i, f_j):
   (a) Test whether f_i dominates f_j, if so, return 1 (full dominance)
   (b) Try to solve the problem numerically using Nelder and Mead and BFGS methods [18]
   (c) Return the highest value found
3. Normalise the dominance degrees according to (31)
4. For each classifier f_i calculate value of $p_{\widetilde{\Delta},\succeq}(f_i)$

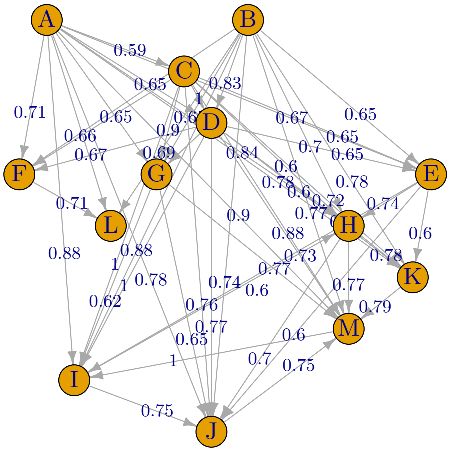


Fig. 4: Graph showing domination degrees ($[\![f_1 \succeq_{1st} f_2]\!]$) obtained with proposed approach for selected classifiers.

More details on evaluation procedure including data pre– and post–processing can be found on GitHub repository.[3]

Graph on Figure 4 presents selected best classifiers with the lowest $p_{\widetilde{\Delta},\succeq}(f_i)$ value. Significant domination degrees (≥ 0.5) are depicted as arrows pointing from dominating to dominated element. There are 3 classifiers that are not significantly dominated. Those models should be considered as potential candidates for choosing. In Table 6 one can see that those models have the lowest and very similar values of $p_{\widetilde{\Delta},\succeq}(f_i)$ which should be used as a final criteria for model selection.

One can see that all comparison methods gave similar results. However, there are few interesting cases which will be discussed here. First one is model J, the best model according to statistical preference. However, this is not confirmed by other methods. The reason of such behaviour is that statistical preference based method, in contrast to other ones, does not take into account the difference in weight between false negatives and false positives. Therefore its results are more similar to those obtained with accuracy.

Second interesting case concerns classifiers D and E. D is being dominated according to First Stochastic Dominance while E is not. This contrasts with the fact that D performs better on other performance measures listed in Table 6 (except sensitivity). Such situation occurred because model C classifies all patients exactly the same as D, except one for which it gives better response (C dominates D, see Fig. 2a). One may say that C is a strictly better version of D. Therefore D should not be chosen as the best classifier. However, this is not true for model E so it still may be considered as a the candidate.

[3] https://github.com/bikol/stochastic-orders-evaluation

Table 6: Summary of various evaluation methods. Shortcuts in header stand for: DEC – Decisiveness, ACC – Accuracy, SEN – Sensitivity, SPC – Specificity.

Model	DEC	ACC	SEN	SPC	Cost	$\succeq_{1st}$	$\succeq_{sp}$	$p_{\widetilde{\Delta},\succeq_{1st}}$
A	0.949	0.886	0.902	0.878	70	0	0.961	0.567
B	0.966	0.876	0.902	0.864	72	0	1.051	0.571
C	0.971	0.876	0.900	0.867	72	0	5.401	0.592
D	0.971	0.871	0.900	0.858	74.5	1	2.293	1.000
E	0.971	0.865	0.918	0.843	75.5	0	2.176	0.675
F	0.931	0.877	0.917	0.861	76	0	0.773	0.713
G	1.000	0.857	0.885	0.846	77.5	0	3.159	0.904
H	0.943	0.885	0.857	0.897	78	1	2.796	0.835
I	0.971	0.859	0.900	0.842	79.5	1	0.727	1.000
J	0.920	0.901	0.826	0.930	80	0	7.462	0.815
K	0.920	0.894	0.848	0.913	80	0	3.186	0.814
L	0.874	0.895	0.909	0.890	80	0	0.583	0.706
M	1.000	0.851	0.731	0.902	100	1	2.739	0.900

5 Discussion and further work

This paper presents an approach to applying stochastic orderings to evaluate classification algorithms for low quality data. We discussed some known stochastic orderings along with practical notes about their application to medical diagnosis support problem. The difficulties that have arisen were our motivation to propose new approach based on fuzzy cost function. The new method allows to compare any two classifiers, but does not require precise definition of the cost function.

All proposed methods were evaluated on real life medical data that comes from recent study on application of aggregation operators to supporting ovarian tumor diagnosis [2]. We were able to obtain results very similar to those previously reported but adopting much weaker assumptions about costs values. This is especially important in this specific problem because as there are still no reliable information on how to estimate costs in medical diagnostics.

Our proposed approach allows to associate numerical metric to each classifier (similarly as in total cost method). This is very useful as it enables the use of this method in more complex evaluation and learning procedures such as cross validation.

As future research we want to evaluate the stability of domination degrees while we slightly change fuzzy cost values. Such stability is very problematic in classical total cost method, where even small changes in costs may lead to big changes in obtained classifier order. As a second line of further research we want to investigate other approaches to fuzzify First Stochastic Dominance based classifier evaluation method such as application of linguistic quantification.

References

1. P. Żywica, A. Wójtowicz, et al., Improving medical decisions under incomplete data using interval–valued fuzzy aggregation, in: Proceedings of 9th European Society for Fuzzy Logic and Technology (EUSFLAT), Gijón, Spain, 2015, pp. 577–584.
2. A. Wójtowicz, P. Żywica, et al., Solving the problem of incomplete data in medical diagnosis via interval modeling, Applied Soft Computing 47 (2016) 424–437.
3. M. Diering, K. Dyczkowski, A. Hamrol, New method for assessment of raters agreement based on fuzzy similarity, in: 10th International Conference on Soft Computing Models in Industrial and Environmental Applications, Springer, 2015, pp. 415–425.
4. A. Stachowiak, K. Dyczkowski, A similarity measure with uncertainty for incompletely known fuzzy sets, in: IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint, IEEE, 2013, pp. 390–394.
5. M. Stukan, M. Dudziak, et al., Usefulness of diagnostic indices comprising clinical, sonographic, and biomarker data for discriminating benign from malignant ovarian masses, Journal of Ultrasound in Medicine 34 (2) (2015) 207–217.
6. R. Moszyński, P. Żywica, et al., Menopausal status strongly influences the utility of predictive models in differential diagnosis of ovarian tumors: An external validation of selected diagnostic tools, Ginekologia Polska 85 (12) (2014) 892–899.
7. N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, New York, 2011.
8. S. Hatch, Snowball in a Blizzard: A Physician's Notes on Uncertainty in Medicine, Basic Books, New York, 2016.
9. I. Couso, L. Sánchez, Generalized stochastic orderings applied to the study of performance of machine learning algorithms for low quality data, in: Proceedings of 16th International Fuzzy Systems Association World Congress and 9th Conference of the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT 2015), Atlantis Press, 2015, pp. 1534–1541.
10. I. Couso, L. Sánchez, Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach, Information Sciences 358 (2016) 129–150.
11. M. Shaked, G. Shanthikumar, Stochastic orders, Springer Science & Business Media, 2007.
12. L. J. Savage, The foundations of statistics, Courier Corporation, 1972.
13. J. Hadar, W. R. Russell, Rules for ordering uncertain prospects, The American Economic Review 59 (1) (1969) 25–34.
14. H. A. David, The method of paired comparisons, Vol. 12, Charles Griffin & D. Ltd., London, 1963.
15. L. Page, S. Brin, et al., The PageRank citation ranking: Bringing order to the web, Tech. rep., Stanford InfoLab (1999).
16. I. Couso, D. Dubois, A perspective on the extension of stochastic orderings to fuzzy random variables, in: Proceedings of 16th International Fuzzy Systems Association World Congress and 9th Conference of the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT 2015), Atlantis Press, 2015, pp. 1486–1492.
17. D. Dubois, H. Prade, Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets, Fuzzy Sets and Systems 192 (2012) 3–24.
18. S. Wright, J. Nocedal, Numerical optimization, Springer Science 35 (1999) 67–68.