A Pairwise Label Ranking Method with Imprecise Scores and Partial Predictions

S. Destercke

Université de Technologie de Compiegne U.M.R. C.N.R.S. 7253 Heudiasyc Centre de
recherches de Royallieu F-60205 Compiegne Cedex FRANCE
Tel: +33 (0)3 44 23 79 85
Fax: +33 (0)3 44 23 44 77
sebastien.destercke@hds.utc.fr

Abstract. In this paper, we are interested in the label ranking problem. We are more specifically interested in the recent trend consisting in predicting partial but more accurate (i.e., making less incorrect statements) orders rather than complete ones. To do so, we propose a ranking method based on pairwise imprecise scores obtained from likelihood functions. We discuss how such imprecise scores can be aggregated to produce interval orders, which are specific types of partial orders. We then analyse the performances of the method as well as its sensitivity to missing data and parameter values.

Key words: Label ranking, imprecise probabilities, Pairwise voting

1 Introduction

In recent years, learning problems with structured outputs have received a growing interest. Such problems appear in a variety of applications fields requiring to deal with complex data: natural language treatment [6], biological data [32], image analysis…

In this paper, we are concerned with the problem of *label ranking*, where one has to learn a mapping from instances to rankings (complete orders) defined over a finite number of labels. Different methods have been proposed to perform this task. Ranking by pairwise comparison (RPC) [25] transforms the problem of label ranking into binary classification problems, combining all results to obtain the final ranking. Constraint classification and log-linear models [23,15] intend to learn, for each label, a (linear) utility function from which the ranking is deduced. Other approaches propose to fit a probabilistic ranking model (Mallows, Placket-Luce [28]) using different approaches (instance-based, linear models, etc. [29,10]).

Recently, some authors [13] have discussed the interest, in label ranking and more generally in preference learning problems, to predict partial orders rather than complete rankings. Such an approach can be seen as an extension of the reject option implemented in learning problems [3] or of the fact of making partial predictions [14]. Such cautious predictions can prevent harmful decisions based on incorrect predictions. In practice, current methods [13] consist in thresholding a pairwise comparison matrix containing probabilistic estimates. More recently, it was shown [12] that probabilities issued from Placket-Luce and Mallows models are particularly interesting in such a thresholding approach, as they are guaranteed to produce consistent orders (i.e., without cycles) that belong to the family of semi-orders.

In this paper, we adopt a different approach in which we propose to use imprecise probabilities and non-parametric estimations to predict partial orderings. As making partial predictions is one central feature of imprecise probabilistic approaches [14], it seems interesting to investigate how one can use such approaches to predict partial orders. In addition, these approaches are also well-designed to cope with the problem of missing or incomplete data [33].

More precisely, we extend the proposal of [25] to imprecise estimates, and propose to get these estimates from a method based on instance-based learning and likelihood functions. The paper is organized as follows: Section 2 discusses the basics of label ranking and of label ranking evaluation when predicting complete and partial orders. Section 3 then presents the method. It first provides a means to obtain imprecise estimates from a likelihood-based approach, before discussing how such estimates can be aggregated to produce interval orders (a sub-family of partial orders including semi-orders) as predictions. Finally, Section 4 ends up with experiments performed on various data sets.

2 Preliminaries

This section introduces the necessary elements concerning label ranking problems.

2.1 Label Ranking Problem

The usual goal of classification problems is to associate an instance $\mathbf{x}$ coming from an instance space $\mathscr{X}$ to a single (preferred) label of the space $\Lambda = \{\lambda_1, \dots, \lambda_m\}$ of possible classes. Label ranking problems correspond to the case where an instance $\mathbf{x}$ is no longer associated to a single label of $\mathscr{X}$ but to a total order over the labels, that is a complete, transitive, and asymmetric relation $\succ_{\mathbf{x}}$ over $\Lambda \times \Lambda$, or equivalently to a complete ranking over the labels $\Lambda = \{\lambda_1, \dots, \lambda_m\}$. Hence, the space of prediction is now the whole set $\mathscr{L}(\Lambda)$ of complete rankings of Λ. It is equivalent to the set of permutations of Λ and contains $|\mathscr{L}(\Lambda)| = m!$ elements. We can identify a ranking $\succ_{\mathbf{x}}$ with a permutation $\sigma_{\mathbf{x}}$ on $\{1, \dots, m\}$ such that $\sigma_{\mathbf{x}}(i) < \sigma_{\mathbf{x}}(j)$ iff $\lambda_i \succ_{\mathbf{x}} \lambda_j$, as they are in one-to-one correspondence. In the following, we will use the terms rankings and permutations interchangeably.

The task in label ranking is the same as the task in usual classification: to use the training instances $(\mathbf{x}_i, y_i)$, $i = 1, \dots, n$ to estimate the theoretical conditional probability measure $P_{\mathbf{x}} : 2^{\mathscr{L}(\Lambda)} \to [0, 1]$ associated to an instance $\mathbf{x} \in \mathscr{X}$. Ideally, observed outputs y_i should be complete orders over Λ, however this is seldom the case and in this paper we allow training instance outputs y_i to be incomplete (i.e., partial orders over Λ).

In label ranking problems the size of the prediction space quickly increases, even when Λ is of limited size (for instance, $|\mathscr{L}(\Lambda)| = 3628800$ for $m = 10$). This makes the estimation of $P_{\mathbf{x}}$ difficult and potentially quite inaccurate if only little data is available, hence an increased interest in providing accurate yet possibly partial predictions. This rapid increase in $|\mathscr{L}(\Lambda)|$ size also means that estimating directly the whole measure $P_{\mathbf{x}}$

is in general untractable except for very small problems. The most usual means to solve this issue is either to decompose the problem into many simpler ones or to assume that $P_{\mathbf{x}}$ follows some parametric law. In this paper, we shall focus on the pairwise decomposition approach, recalled and extended in Section 3. To simplify notations, we will drop the subscript $_{\mathbf{x}}$ in the following when there is no possible ambiguity.

2.2 Evaluating Partial Predictions

The classical task of label ranking is to predict a complete ranking $\hat{y} \in \mathscr{L}(\Lambda)$ of labels as close as possible to an observed complete ranking y. When the observed and predicted rankings y and $\hat{y}$ are complete, various accuracy measures [25] (0/1 accuracy, Spearman's rank, ...) have been proposed to measure how close $\hat{y}$ is to y. In this paper, we retain Kendall's Tau, as it will be generalized to measure the quality of partial predictions. Given y and $\hat{y}$, Kendall's Tau is

$$A_\tau(y,\hat{y}) = \frac{C-D}{m(m-1)/2} \tag{1}$$

where $C = |\{(\lambda_i,\lambda_j)|(\sigma(i) < \sigma(j)) \wedge (\hat{\sigma}(i) < \hat{\sigma}(j))\}|$ is the number of concording pairs of labels in the two rankings, and $D = |\{(\lambda_i,\lambda_j)|(\sigma(j) < \sigma(i)) \wedge (\hat{\sigma}(i) < \hat{\sigma}(j)))\}|$ the number of discording pairs of labels. $A_\tau(y,\hat{y})$ has value 1 when $y = \hat{y}$ and -1 when $\hat{y}$ and y are reversed rankings.

$A_\tau(y,\hat{y})$ assumes that the prediction $\hat{y}$ is a complete ranking and that the model can compare each pair of labels in a reliable way. Such an assumption is quite strong, especially if the information in the training samples is not complete (e.g., incomplete rankings). When we allow the prediction $\hat{y}$ to be a partial order, $A_\tau(y,\hat{y})$ needs to be adapted.

[13] propose to decompose the usual accuracy measures into two components: the correctness (CR) measuring the quality of the predicted comparisons; and the completeness (CP) measuring the completeness of the prediction. They are defined as

$$CR(y,\hat{y}) = \frac{C-D}{C+D} \quad \text{and} \quad CP(y,\hat{y}) = \frac{C+D}{m(m-1)/2}, \tag{2}$$

where $\hat{y}$ is a partial order and where C and D have the same definitions as in Eq. (1). When the predicted order $\hat{y}$ is complete ($C + D = m(m-1)/2$), $CR(y,\hat{y}) = A_\tau(y,\hat{y})$ and $CP(y,\hat{y}) = 1$, while $CP(y,\hat{y}) = 0$ and by convention $CR(y,\hat{y}) = 1$ if no comparison is done (as all orders are then considered possible).

To summarize, the following assumptions are made in this paper:

- the theoretical model we seek to estimate is a probability measure P defined on the space $\mathscr{L}(\Lambda)$ of complete rankings;
- training instance outputs $y_i, i = 1,\ldots,n$ are allowed to be incompletely observed (i.e., partial orders), while test instances are assumed to be fully observed (i.e. complete rankings);
- the predictions $\hat{y}$ are allowed to be partial orders.

3 Partial Orders Prediction Method

This section describes our likelihood pairwise comparison (LPC) method. It first recalls the principle of pairwise decomposition (Section 3.1). It then details the proposed likelihood-based method used to obtain imprecise estimates (Section 3.2) before discussing how such estimates can be aggregated to obtain partial orders (Section 3.3).

3.1 Pairwise Decomposition

Pairwise decomposition is a well-known procedure used in classification to simplify the initial problem [27,20], which is divided into several binary problems then combined into a final prediction. A similar approach can also be used in preference learning and label ranking problems: it consists in estimating, for each pair of labels λ_i, λ_j, the probabilities $P(\{\lambda_i \succ \lambda_j\})$ or $P(\lambda_j \succ \lambda_i)$ and then to predict an (partial) order on Λ from such estimates. In practice, this can be done by decomposing the data set into ${}^{(m-1)m}/_{2}$ data sets, one for each pair. This decomposition is illustrated in Figure 1 on an imaginary label ranking training data set with four input attributes.

X_1	X_2	X_3	X_4	Y
107.1	25	*Blue*	60	$\lambda_1 \succ \lambda_3 \succ \lambda_2$
−50	10	*Red*	40	$\lambda_2 \succ \lambda_3$
200.6	30	*Blue*	58	$\lambda_2 \succ \lambda_1 \succ \lambda_3$
107.1	5	*Green*	33	$\lambda_1 \succ \lambda_2$
…	…	…	…	…

pair λ_1, λ_2

X_1	…	X_4	Y
107.1		60	$\lambda_1 \succ \lambda_2$
200.6		58	$\lambda_2 \succ \lambda_1$
107.1		33	$\lambda_1 \succ \lambda_2$
…		…	…

pair λ_1, λ_3

X_1	…	X_4	Y
107.1		60	$\lambda_1 \succ \lambda_3$
200.6		58	$\lambda_1 \succ \lambda_3$
…		…	…

pair λ_2, λ_3

X_1	…	X_4	Y
107.1		60	$\lambda_3 \succ \lambda_2$
−50		40	$\lambda_2 \succ \lambda_3$
200.6		58	$\lambda_2 \succ \lambda_3$
…		…	…

Fig. 1. Pairwise decomposition of rankings

From a data perspective, working with pairwise comparisons is not very restrictive, as almost all models working with preferences can be decomposed into such pairwise preferences: complete rankings, top-t preferences where only the first t labels $\sigma(1), \ldots, \sigma(t)$ are ordered [8], rankings of subsets A of Λ where only labels in A are ordered [22], rankings over a partition of Λ (or bucket orders) [21]. Pairwise preferences are even more general, as most of the previous models cannot model a unique preference $\lambda_i \succ \lambda_j$ [7].

For each pair λ_i, λ_j, the data $(\mathbf{x}_k, y_k)$ is retained if it contains the information $\lambda_i \succ \lambda_j$ or $\lambda_j \succ \lambda_i$, and is forgotten otherwise. Using all data for which λ_i, λ_j are compared, the goal is then to estimate the probability

$$P(\{\lambda_i \succ \lambda_j\}) = 1 - P(\{\lambda_j \succ \lambda_i\}). \tag{3}$$

Any probabilistic binary classifier can be used to estimate this probability once the data set has been decomposed (possibly by mapping $\lambda_i \succ \lambda_j$ to value 1 and $\lambda_i \prec \lambda_j$ to 0 or -1). We will denote by $\hat{P}(\{\lambda_i \succ \lambda_j\})$ the obtained estimate of the theoretical probability $P(\{\lambda_i \succ \lambda_j\})$.

Rather than using only a precise estimate $\hat{P}(\{\lambda_i \succ \lambda_j\})$ as score, we propose in this paper to learn an imprecise estimate of $P(\{\lambda_i \succ \lambda_j\})$ in the form of an interval $[\hat{P}(\{\lambda_i \succ \lambda_j\})] = [\underline{\hat{P}}(\{\lambda_i \succ \lambda_j\}), \overline{\hat{P}}(\{\lambda_i \succ \lambda_j\})]$ and to use such imprecise estimates to predict partial orders. The next section introduces a method inspired from imprecise probabilistic approaches that provides a continuous range of nested imprecise estimates, going from a precise one ($\underline{\hat{P}}(\{\lambda_i \succ \lambda_j\}) = \overline{\hat{P}}(\{\lambda_i \succ \lambda_j\})$) to a completely imprecise one ($[\hat{P}(\{\lambda_i \succ \lambda_j\})] = [0,1]$).

3.2 Pairwise Imprecise Estimates by Contour Likelihood

One of the goal of imprecise probabilistic methods is to extend classical estimation methods to provide imprecise but cautious estimates of quantities. Some of them extend Bayesian approaches by considering sets of priors [4], while others extend maximum likelihood approaches [9]. One of the latter is retained here, called contour likelihood method [9], as it will allow us to go from a totally precise to a totally imprecise estimate in a smooth way.

Given a parameter θ taking its values on a space Θ (e.g., $P(\{\lambda_i \succ \lambda_j\})$ on $[0,1]$) and a positive likelihood function $L : \Theta \to \mathbb{R}^+$, we call *contour likelihood* L^* the function

$$L^*(\theta) = \frac{L(\theta)}{\max_{\theta \in \Theta} L(\theta)}. \tag{4}$$

Using this function as an imprecise probabilistic model (and more specifically as a possibility distribution) has been justified by different authors [17,31,9], and we refer to them for a thorough discussion. Historically, the use of relative likelihood to get estimates of parameters dates back to Fisher [24] and Birnbaum [5].

Imprecise estimates are then obtained by using the notion of *β-cut*. Given a value $\beta \in [0,1]$, the β-cut L^*_β of L^* is the set such that

$$L^*_\beta = \{\theta \in \Theta | L^*(\theta) \geq \beta\}. \tag{5}$$

Given Eq. (4), we have $L^*_1 = \arg\max_{\theta \in \Theta} L(\theta)$ (the precise maximum likelihood estimator) and $L^*_0 = \Theta$ (the whole set of possible parameter values). In between, we have that $L^*_{\beta_1} \subseteq L^*_{\beta_2}$ for any values $\beta_1 > \beta_2$, that is the lower the value of β, the more imprecise and cautious is our estimate L^*_β. Such estimates are usually simple to obtain and have the advantage (e.g., over frequentist confidence intervals) to follow the likelihood principle, that is to say they depend on the sampling model and data only through the likelihood function (they do not require extra information such as prior probabilities).

In a binary space where $\theta \in [0,1]$ is the probability of success, Eq. (4) becomes

$$L^*(\theta) = \frac{\theta^s(1-\theta)^{n-s}}{(s/n)^s(1-s/n)^{n-s}} \tag{6}$$

with n the number of observations, s the number of success and $\arg\max_{\theta\in\Theta} L(\theta) = s/n$.

Once they are decomposed into pairwise preferences, we can use training examples $(\mathbf{x}_k, y_k)$, $k = 1, \ldots, n$ and Eq. (6) to estimate $P(\{\lambda_i \succ \lambda_j\})$ for any pair λ_i, λ_j and for a new instance $\mathbf{x}$. To do so, we assume that a metric d is defined (or can be defined) on $\mathscr{X}$ and we propose a simple instance-based strategy.

For a given upper distance $\overline{d}$, let $\mathscr{N}_{\overline{d},\mathbf{x}} = \{\mathbf{x}_i : d(\mathbf{x}, \mathbf{x}_i) \leq \overline{d}\}$ the set of training instances whose distance from $\mathbf{x}$ is lower than some (upper) distance $\overline{d}$. In this set of training examples, let us denote by

- $\mathscr{N}_{\overline{d},\mathbf{x}}(i,j) = \{\mathbf{x}_k : \mathbf{x}_k \in \mathscr{N}_{\overline{d},\mathbf{x}}, (\lambda_i \succ \lambda_j) \vee (\lambda_j \succ \lambda_i) \in y_k\}$ the set of all instances that provides a comparison for the labels λ_i and λ_j;
- $\mathscr{N}_{\overline{d},\mathbf{x}}(i > j) = \{\mathbf{x}_k : \mathbf{x}_k \in \mathscr{N}_{\overline{d},\mathbf{x}}(i,j) \wedge (\lambda_i \succ \lambda_j) \in y_k\}$ the set of items where λ_i is preferred to λ_j.

Using these information, interval estimates $[\hat{P}(\{\lambda_i \succ \lambda_j\})]_\beta$ are then simply obtained using Eq. (6) with $|\mathscr{N}_{\overline{d},\mathbf{x}}(i,j)|$ the number of observations, $|\mathscr{N}_{\overline{d},\mathbf{x}}(i > j)|$ the number of successes and β a fixed level of confidence.

Figure 2 pictures two contour likelihoods together with estimates obtained for a given β. As can be seen from the picture, for a given β the imprecision of the estimate $[\hat{P}(\{\lambda_i \succ \lambda_j\})]_\beta$ will depend on the amount of data used to compute $L^*(\theta)$.

As other instance-based methods (e.g., k-nearest neighbour), this approach is based on the assumption that $P_\mathbf{x}$ is *constant* around the instance $\mathbf{x}$. In practice, this means that $\overline{d}$ should not be too high, but also not too small (otherwise there may be no data in the neighbourhood). As we fix $\overline{d}$ rather than the number of neighbours, the presence of missing data will lead to a lower value of $|\mathscr{N}_{\overline{d},\mathbf{x}}(i,j)|$ and to a more imprecise estimate (for a given β). The amount of missing data is therefore automatically considered by the method.

Once precise or imprecise estimates $[\hat{P}(\{\lambda_i \succ \lambda_j\})]$ are obtained, the next step is to combine them to obtain a predicted (partial) order $\hat{y}$. In this paper, we extend the voting approach detailed in [25]. Note that the values that (theoretical) probabilities

$$P(\{\lambda_i \succ \lambda_j\}) = \sum_{y \in \mathscr{L}(\Lambda), \lambda_i \succ \lambda_j} P(\{y\})$$

can assume are constrained, as they are linked by the following weak transitivity relation for any three labels $\lambda_i, \lambda_j, \lambda_k$ [25]:

$$P(\{\lambda_i \succ \lambda_j\}) \geq P(\{\lambda_i \succ \lambda_k\}) + P(\{\lambda_k \succ \lambda_j\}) - 1. \tag{7}$$

This relation may not be satisfied by all estimates $[\hat{P}(\{\lambda_i \succ \lambda_j\})]$. Luckily, post-processing methods of estimates $[\hat{P}(\{\lambda_i \succ \lambda_j\})]$ (such as the voting approach) usually do not require this relation to be satisfied to predict a consistent order.

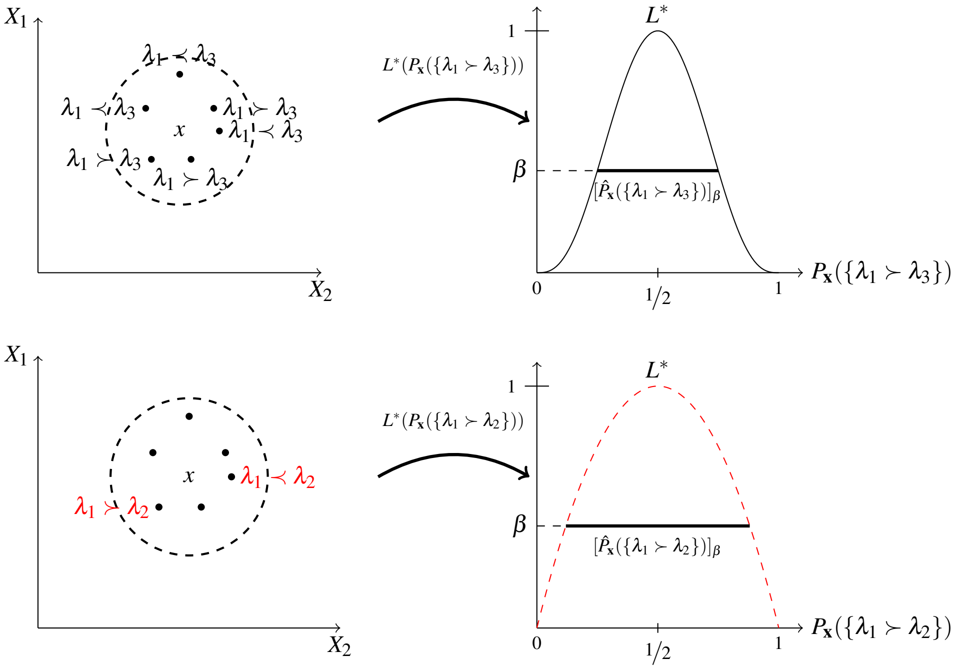

X_1
$\lambda_1 \prec \lambda_3$
$\lambda_1 \prec \lambda_3$
$\lambda_1 \succ \lambda_3$
x
$\lambda_1 \prec \lambda_3$
$\lambda_1 \succ \lambda_3$
$\lambda_1 \succ \lambda_3$
X_2
$L^*(P_{\mathbf{x}}(\{\lambda_1 \succ \lambda_3\}))$
1
L^*
β
$[\hat{P}_{\mathbf{x}}(\{\lambda_1 \succ \lambda_3\})]_\beta$
0
1/2
1
$P_{\mathbf{x}}(\{\lambda_1 \succ \lambda_3\})$
X_1
x
$\lambda_1 \prec \lambda_2$
$\lambda_1 \succ \lambda_2$
X_2
$L^*(P_{\mathbf{x}}(\{\lambda_1 \succ \lambda_2\}))$
1
L^*
β
$[\hat{P}_{\mathbf{x}}(\{\lambda_1 \succ \lambda_2\})]_\beta$
0
1/2
1
$P_{\mathbf{x}}(\{\lambda_1 \succ \lambda_2\})$

Fig. 2. Imprecise estimates through β-cut of relative likelihood: illustration

3.3 Aggregating Imprecise Estimates to Get Partial Orders

In [25], precise estimates $\hat{P}(\{\lambda_i \succ \lambda_j\})$ are considered as (weighted) votes and aggregated for each label λ_i into a global score

$$\hat{S}(i) = \sum_{j \in \{1,\ldots,m\} \setminus i} \hat{P}(\{\lambda_i \succ \lambda_j\}). \tag{8}$$

Labels are then ordered according to their scores $\hat{S}(i)$, that is $\lambda_i \succ \lambda_j$ if and only if $\hat{S}(i) \geq \hat{S}(j)$. It has been shown [25] that using this strategy provides optimal predictions for the Spearman rank correlation in the sense that it maximizes its expected accuracy if $\hat{P}(\{\lambda_i \succ \lambda_j\}) = P(\{\lambda_i \succ \lambda_j\})$ (Kendall Tau can also be optimized by using only $P(\{\lambda_i \succ \lambda_j\})$, however it requires to solve the NP-hard minimum feedback arc set problem [1]). We will denote by $S(j)$ the (theoretical) scores that would have been obtained by using the theoretical measure P.

Example 1. We consider the space of labels $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ with the following matrix of estimates $\hat{P}(\{\lambda_i \succ \lambda_j\})$ and scores $\hat{S}(i)$

$\hat{P}(\{\lambda_i \succ \lambda_j\})$	λ_1	λ_2	λ_3	$\hat{S}(i)$
λ_1		0.6	0.6	1.2
λ_2	0.4		0.3	0.7
λ_3	0.4	0.7		1.1

The obtained prediction is $\lambda_2 \prec \lambda_3 \prec \lambda_1$ (Note that estimates $\hat{P}(\{\lambda_i \succ \lambda_j\})$ satisfy constraints (7) and could originate from a theoretical model P).

Let us now deal with the case where estimates $[\hat{P}(\{\lambda_i \succ \lambda_j\})]$ are imprecise (here, originating from the method presented in Section 3.2). It is straightforward to extend Eq. (8) to imprecise estimates by defining the imprecise score $[\hat{S}(i)]$ as

$$\begin{aligned} [\hat{S}(i)] &= \sum_{j \in \{1,\ldots,m\} \setminus i} [\hat{P}(\{\lambda_i \succ \lambda_j\})] \\ &= [\sum_{j \in \{1,\ldots,m\} \setminus i} \underline{\hat{P}}(\{\lambda_i \succ \lambda_j\}), \sum_{j \in \{1,\ldots,m\} \setminus i} \overline{\hat{P}}(\{\lambda_i \succ \lambda_j\})]. \end{aligned} \tag{9}$$

There are multiple ways to compare intervals $[\hat{S}(i)]$ to get a partial or a complete order. Let us denote $\hat{y}_D$ the prediction obtained by a decision rule D and intervals $[\hat{S}(i)]$. We think that decision rules producing partial orders from interval-valued scores should obey at least the two following properties:

Definition 1 (Imprecision monotonicity). *Consider two assessments $[\hat{P}]$ and $[\hat{P}^*]$ such that, for every pair $i,j \in \{1,\ldots,m\}$ we have $[\hat{P}(\{\lambda_i \succ \lambda_j\})] \subseteq [\hat{P}^*(\{\lambda_i \succ \lambda_j\})]$. A decision rule D is said* imprecision monotonic *if*

$$\lambda_i \succ \lambda_j \in \hat{y}_D^* \Rightarrow \lambda_i \succ \lambda_j \in \hat{y}_D$$

for any pair i,j and with $\hat{y}_D, \hat{y}_D^*$ the predictions produced using estimates $[\hat{P}]$ and $[\hat{P}^*]$, respectively.

This property basically says that getting less information cannot make our prediction more precise, in the sense that every label pair comparable according to estimates $[\hat{P}^*]$ should also be comparable according to $[\hat{P}]$ under decision rule D. If we denote by $\mathscr{C}(\hat{y})$ the set of linear extensions (i.e., of completions into complete orders) of the partial order $\hat{y}$, another way to formalise imprecision monotonicity is to ask $\mathscr{C}(\hat{y}_D) \subseteq \mathscr{C}(\hat{y}_D^*)$, that is to require every possible completion of $\hat{y}_D$ to be also a completion of $\hat{y}_D^*$. The second property is not related to imprecision, but to the coherence between the predicted partial order and the complete order that would be obtained using the theoretical model P.

Definition 2 (Model coherence). *Let $S(j)$ be the theoretical scores, y the associated complete order and $[\hat{P}]$ an assessment with associated scores $[\hat{S}(j)]$. Then, a decision rule D is said* model coherent *if*

$$S(j) \in [\hat{S}(j)] \forall j \in \{1,\ldots,m\} \Rightarrow y \in \mathscr{C}(\hat{y}_D)$$

This property requires that if our estimates are consistent with the theoretical model (i.e., include the true value), then the optimal complete ranking is an extension of our prediction. That is, our prediction is totally consistent with the optimal solution, but is possibly incomplete. In particular, satisfying model coherence ensures that the prediction optimizing Spearman rank correlation is in $\mathscr{C}(\hat{y}_D)$, provided $P \in [\hat{P}]$.

To produce a partial ranking from intervals $[\hat{S}(j)]$, we propose to use the following decision rule, that we call *strict dominance* and denote $\mathscr{I}$:

$$\lambda_i \succ_{\mathscr{I}} \lambda_j \Leftrightarrow \underline{\hat{S}}(i) \geq \overline{\hat{S}}(j).$$

That is, label λ_i is ranked before label λ_j only when we are certain that the score of λ_j is lower than the one of λ_i. Partial orders obtained following this rule correspond to so-called interval-orders [18], that have been widely studied in the literature. The next proposition shows that such a procedure satisfies the properties we find appealing.

Proposition 1. *The ordering $\succ_{\mathscr{I}}$ is imprecision monotonic and model coherent.*

Proof. Let $\succ_{\mathscr{I}}, \succ^*_{\mathscr{I}}$ be the interval orders obtained by estimates $[\hat{P}]$ and $[\hat{P}^*]$ with rule $\mathscr{I}$.

Imprecision monotonic: if for every pair $i, j \in \{1,\dots,m\}$ we have $[\hat{P}(\{\lambda_i \succ \lambda_j\})] \subseteq [\hat{P}^*(\{\lambda_i \succ \lambda_j\})]$, then $[\hat{S}(i)] \subseteq [\hat{S}^*(i)]$ for any label λ_i and $\lambda_i \succ^*_{\mathscr{I}} \lambda_j$ implies $\lambda_i \succ_{\mathscr{I}} \lambda_j$, since the inequalities

$$\overline{\hat{S}}(j) \leq \overline{\hat{S}^*}(j) \leq \underline{\hat{S}^*}(i) \leq \underline{\hat{S}}(i)$$

hold. The second is due to $\lambda_i \succ^*_{\mathscr{I}} \lambda_j$, while the first and third are due to $[\hat{S}(i)] \subseteq [\hat{S}^*(i)]$ for any label λ_i. This is sufficient to show imprecision monotonicity.

Model coherence: assume that $P(\{\lambda_i \succ \lambda_j\}) \in [\hat{P}(\{\lambda_i \succ \lambda_j\})]$. Then we can show that $\lambda_i \succ_{\mathscr{I}} \lambda_j$ implies $\lambda_i \succ \lambda_j$, where $\succ$ is the ordering obtained from P. Simply observe that inequalities

$$S(j) \leq \overline{\hat{S}}(j) \leq \underline{\hat{S}}(i) \leq S(i)$$

hold as $\lambda_i \succ \lambda_j$ and $S(j) \in [\hat{S}(j)] \forall j \in \{1,\dots,m\}$ by definition. This is sufficient to show model coherence.

Example 2. Consider the following matrix of imprecise scores that include the matrix of Example 1

$\hat{P}(\{\lambda_i \succ \lambda_j\})$	λ_1	λ_2	λ_3	$S(i)$
λ_1		$[0.4, 0.6]$	0.6	$[1, 1.2]$
λ_2	$[0.4, 0.6]$		$[0.1, 0.3]$	$[0.5, 0.9]$
λ_3	0.4	$[0.7, 0.9]$		$[1.1, 1.3]$

Applying the $\mathscr{I}$ rule results in $\lambda_2 \prec_{\mathscr{I}} \lambda_3$ and $\lambda_2 \prec_{\mathscr{I}} \lambda_1$, without being able to compare λ_3 and λ_1. This prediction is more cautious but coherent with the ordering obtained in Example 1, which was $\lambda_2 \prec \lambda_3 \prec \lambda_1$.

In summary, the likelihood pairwise comparison (LPC) method consists in the following steps:

1. decompose the training data set $(\mathbf{x}_k, y_k)$, $k = 1,\dots,n$ into pairwise data sets;
2. pick a distance $\overline{d}$ and a level $\beta \in [0,1]$;
3. for each pair λ_i, λ_j, take as estimate the interval $[\hat{P}(\{\lambda_i \succ \lambda_j\})]_\beta$ from L^*;
4. compute $[\hat{S}(j)]$ and use strict dominance to predict an interval order $\succ_{\mathscr{I}}$.

By varying β, we can go smoothly from a precise ordering $\succ_{\mathscr{I}}$ ($\beta = 1$) to an ordering making no comparison at all ($\beta = 0$), similarly to what is done in [13] by varying the threshold. Note, however, that this approach is quite different from [13,12]:

- we rely on aggregation of imprecise estimates to predict partial orders rather than thresholding a precise model (in [13,12], the prediction $\lambda_i \succ \lambda_i$ is made if $\hat{P}_{\mathbf{x}}(\{\lambda_i \succ \lambda_j\}) \geq \alpha$ with $\alpha \in [0.5, 1]$);
- we always predict an interval order (a family of partial orders that includes semi-orders, the type of orders predicted in [12]) and do not have to face issues related to the presence of cycles [13];
- we use a non-parametric estimation method rather than parametric probabilities [12], which makes our approach computationally more demanding. As the aggregation methods presented in Section 3.3 applies to any imprecise estimates, it would be interesting to study how confidence intervals can be extracted from estimated parametric models, or to which extent are the results affected by considering other imprecise estimates (e.g., confidence intervals).

4 Experiments

In this section, we first compare the performances of our approach with two other techniques in the case of complete order predictions. We use the WEKA-LR [2] implementation. We also discuss the behaviour of our approach with respect to missing data.

The datasets used in the experiments come from the UCI machine repository [19] and the Statlog collection [26]. They are synthetic label ranking data sets built either from classification or regression problems. From each original data set, a transformed data set $(\mathbf{x}_i, y_i)$ with complete rankings was obtained by following the procedure described in [11]. A summary of the data sets used in the experiments is given in Table 1.

4.1 Comparative Experiments with Precise Predictions

To show that our approach performs satisfyingly, we apply our method to complete data sets of Table 1 and compare its results with other label ranking approaches in the case where predictions are complete ($\beta = 1$ in Eq. (5)). More precisely, we compare the results of the proposed approach with the Ranking by Pairwise Comparison method (RPC) using a logistic regression as base classifier [25] and the Label Ranking Tree (LRT) method [11]. Note that if $\beta = 1$ in Eq. (5), LPC is equivalent to adopt a ranking by pairwise comparison approach (RPC) with another base classifier.

Kendall tau is used to assess the accuracy of the classifiers, and reported results are averages over a 10-fold cross validation. Concerning the LPC method, the Euclidean distance d was used, with a maximum radius $\overline{d} = a\mathbb{E}_d$ where $a > 0$ multiplies the average distance

$$\mathbb{E}_d = {}^{n(n-1)}/{}_{2} \sum_{\substack{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{x}_1, \ldots, \mathbf{x}_n \\ j \neq i}} d(\mathbf{x}_i, \mathbf{x}_j)$$

between all training instances. As the goal of these experiments is to assess whether our method provides satisfying results, we set $a = 1.0$ (the effect of modifying a when preferences are missing is studied in the next section).

Table 1. Experimental data sets

Data set	Type	#Inst	#Attributes	#Labels
authorship	classification	841	70	4
bodyfat	regression	252	7	7
calhousing	regression	20640	4	4
cpu-small	regression	8192	6	5
elevators	regression	16599	9	9
fried	regression	40768	9	5
glass	classification	214	9	6
housing	regression	506	6	6
iris	classification	150	4	3
pendigits	classification	10992	16	10
segment	classification	2310	18	7
stock	regression	950	5	5
vehicle	classification	846	18	4
vowel	classification	528	10	11
wine	classification	178	13	3
wisconsin	regression	194	16	16

Table 2. Results on precise case

	LPC		RPC		LRT	
Data set	accuracy	rank	accuracy	rank	accuracy	rank
authorship	0.910	1	0.908	2	0.887	3
bodyfat	0.216	2	0.282	1	0.11	3
calhousing	0.273	2	0.244	3	0.357	1
cpu-small	0.421	2	0.449	1	0.423	3
elevators	0.701	3	0.749	2	0.758	1
fried	0.789	3	0.999	1	0.888	2
glass	0.853	3	0.887	2	0.893	1
housing	0.699	2	0.674	3	0.799	1
iris	0.92	2	0.893	3	0.947	1
pendigits	0.879	1	0.932	3	0.942	2
segment	0.880	3	0.934	2	0.956	1
stock	0.792	2	0.779	3	0.892	1
vehicle	0.843	2	0.857	1	0.833	3
vowel	0.805	1	0.652	3	0.795	2
wine	0.947	1	0.914	2	0.88	3
wisconsin	0.451	2	0.634	1	0.328	3
Average rank	1.8		2		2.2	

To compare the different results, Demsar [16] approach is used on the results of Table 2. Friedman test was used on the ranks of algorithm performances for each data-set, finding a value 1.13 for the Chi-square test with 2 degree of freedom and a corresponding p-value of 0.57, hence the null-hypothesis (no significant differences between algorithms) cannot be rejected. The algorithms therefore display comparable performances. It should be noted, however, that no optimisation was performed for the proposed method (either on the value $\overline{d}$ or on the shape of the neighbourhood region).

4.2 Accuracy of Partial Predictions

In the previous section, we have shown that our method is competitive with other label ranking methods in situations where complete orders are predicted ($\beta = 1$ in LPC). In this section, we study the behaviour of LPC with respect to completeness and correctness (2) when we allow for partial predictions, that is $\beta \in [0,1]$ in Eq. (5) and we use the strict dominance rule $\mathscr{I}$ to produce predictions. As in [11,12], we span a whole set of partial orders by going from precise orders ($\beta = 1$) to completely imprecise ones ($\beta = 0$). However we span a richer family of partial orders, namely interval orders.

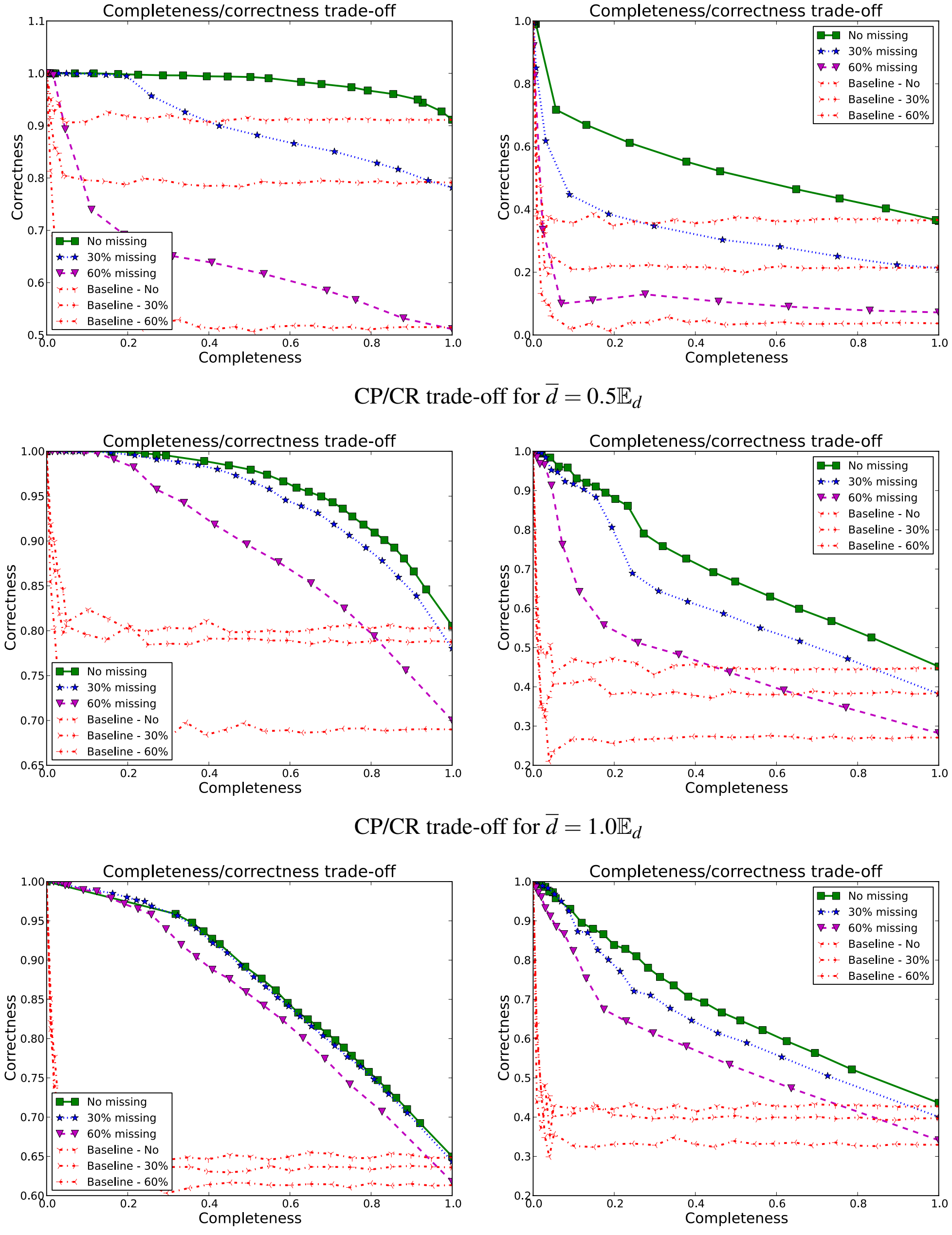
To study how LPC behaves when some pairwise preferences are missing, we also consider incomplete rankings in training instances. Missing preferences in the training data sets are induced with the following strategy [11]: for a given training instance y_k, each label is removed with a probability γ (here, either 30 or 60%).

Intuitively, we may expect the predictions to be more accurate (i.e., predicted comparisons to be more often correct) as they become more partial. That is, as β decreases, the average completeness CP decreases, with the hope that this decrease is counterbalanced by an increase in correctness CR. To verify this intuition, we have compared our approach with the following base-line: for a given β, we have considered the complete ordering obtained with $\beta^* = 1$ in (5), and have randomly removed each pairwise comparison induced by this ordering with a probability $1-\beta$.

Figure 3 shows the evolution of completeness and correctness for two data sets (a classification one, vowel, and a regression one, wisconsin) as β decreases for various choices of $\overline{d}$ and for different percentages of missing data. As expected, the (average) correctness is increasing as completeness decreases for our method, while the baseline that performs random suppression of preferences does not show a significant increase of correctness as completeness decreases. This confirms that our method provides cautious yet more accurate predictions as β decreases.

There are other facts that we may notice from the graphs in Figure 3:

- the higher is the distance $\overline{d}$, the more stable is the evolution of correctness/completeness, showing that LPC with higher distances is less affected by missing preferences. In particular, correctness for a level $\beta = 1$ ($CP = 1$) does not change significantly when $\overline{d}$ is high, whether preferences are missing or not. On the contrary, the effect of missing preferences is quite noticeable for lower values of $\overline{d}$, particularly when β is low. This is not surprising, as a higher $\overline{d}$ means using more training instances to assess the model;
- when there are no missing preference, taking a lower $\overline{d}$ usually provides better correctness than a higher one. This can be explained by the fact that the instance-based



CP/CR trade-off for $\overline{d} = 0.5\mathbb{E}_d$

CP/CR trade-off for $\overline{d} = 1.0\mathbb{E}_d$

CP/CR trade-off for $\overline{d} = 1.5\mathbb{E}_d$

Fig. 3. Results for vowel (left column) and wisconsin (right column) data set

assumption (i.e. assuming $P_{\mathbf{x}}$ constant around $\mathbf{x}$) becomes less and less supported when $\overline{d}$ increases. However, when the number of missing preferences is significant, correctness is usually better for large $\overline{d}$, as the model is then less sensible to such missing preferences.

These experiments suggest that the choice of a good $\overline{d}$ heavily depends on the data: if full rankings are given for each training instance, then $\overline{d}$ should be kept low, while if the information is poor (many partial rankings with few preferences), a higher $\overline{d}$ should be preferred.

5 Conclusion

In this paper, we have introduced the likelihood pairwise comparison (LPC) approach to achieve label ranking by pairwise comparisons, which can be seen as an instance-based non-parametric approach.

Although the method can produce complete rankings as predictions, one of its main interest lies in the ability to produce partial but more accurate orders as predictions. This is done by using the interpretation [17,9] seeing the normalised likelihood function as an imprecise probabilistic model, and more precisely as a possibility distribution from which are derived imprecise estimates. To build this normalised likelihood, we have proposed a simple instance-based approach using the neighbours that are within a given radius of the instance.

Our results indicate that the choice of the distance (i.e., radius) used to estimate our model can be important: a higher distance will usually produce less accurate predictions when preferences are complete and more accurate predictions when preferences are missing, while a lower distance will produce more accurate predictions when preferences are complete, but will be more sensible to missing data.

Compared to [12], our method also guarantees the consistency of predicted partial orders while being potentially more expressive, as predicted partial orders are interval orders (that include semi-orders). First experimental results show a good increase of correctness when partial predictions are considered. In the future, it would be interesting to compare the obtained results to other methods, or to study the problem of predicting partial orders from imprecisely specified parametric models (as non-parametric instance-based methods are computationally costly), possibly combining them with other decision rules of imprecise probabilistic approaches [30].

References

1. Alon, N.: Ranking tournaments. SIAM Journal on Discrete Mathematics 20, 137–142 (2006)
2. Balz, A., Senge, R.: Weka-lr: A label ranking extension for weka (Jun 2011), `http://www.uni-marburg.de/fb12/kebi/research/software/WEKA-LR-PAGE`
3. Bartlett, P., Wegkamp, M.: Classification with a reject option using a hinge loss. The Journal of Machine Learning Research 9, 1823–1840 (2008)
4. Bernard, J.: An introduction to the imprecise dirichlet model for multinomial data. International Journal of Approximate Reasoning 39(2), 123–150 (2005)

5. Birnbaum, A.: On the foundations of statistical inference. Journal of the American Statistical Association 57(298), 269–306 (1962)
6. Bordes, A., Glorot, X., Weston, J., Bengio, Y.: Joint learning of words and meaning representations for open-text semantic parsing. Journal of Machine Learning Research - Proceedings Track 22, 127–135 (2012)
7. Boutilier, C., Lu, T.: Learning mallows models with pairwise preferences. In: Proceedings of the 28th Annual International Conference on Machine Learning - ICML. pp. 145–152 (2011)
8. Busse, L., Orbanz, P., Buhmann, J.: Cluster analysis of heterogeneous rank data. In: ACM International Conference Proceeding Series. vol. 227, pp. 113–120 (2007)
9. Cattaneo, M.: Statistical Decisions Based Directly on the Likelihood Function. Ph.D. thesis, ETH Zurich (2007)
10. Cheng, W., Dembczynski, K., Hüllermeier, E.: Label ranking methods based on the plackett-luce model. In: Proceedings of the 27th Annual International Conference on Machine Learning - ICML. pp. 215–222 (2010)
11. Cheng, W., Hühn, J., Hüllermeier, E.: Decision tree and instance-based learning for label ranking. In: Proceedings of the 26th Annual International Conference on Machine Learning - ICML' 09 (2009)
12. Cheng, W., Hüllermeier, E., Waegeman, W., Welker, V.: Label ranking with partial abstention based on thresholded probabilistic models. In: Advances in Neural Information Processing Systems 25 (NIPS-12). pp. 2510–2518 (2012)
13. Cheng, W., Rademaker, M., De Baets, B., Hüllermeier, E.: Predicting partial orders: ranking with abstention. Machine Learning and Knowledge Discovery in Databases pp. 215–230 (2010)
14. Corani, G., Antonucci, A., Zaffalon, M.: Bayesian networks with imprecise probabilities: Theory and application to classification. Data Mining: Foundations and Intelligent Paradigms pp. 49–93 (2012)
15. Dekel, O., Manning, C.D., Singer, Y.: Log-linear models for label ranking. In: Advances in Neural Information Processing Systems (2003)
16. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 9(7), 1–30 (2006)
17. Dubois, D., Moral, S., Prade, H.: A semantics for possibility theory based on likelihoods. Journal of Mathematical Analysis and Applications 205, 359–380 (1997)
18. Fishburn, P.: Interval Orderings. Wiley (1987)
19. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
20. Fürnkranz, J.: Pairwise classification as an ensemble technique. Machine Learning: ECML 2002 pp. 9–38 (2002)
21. Gionis, A., Mannila, H., Puolamäki, K., Ukkonen, A.: Algorithms for discovering bucket orders from data. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 561–566. ACM (2006)
22. Guiver, J., Snelson, E.: Bayesian inference for plackett-luce ranking models. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 377–384. ACM (2009)
23. Har-peled, S., Roth, D., Zimak, D.: Constraint classification : A new approach to multiclass classification and ranking. In: Advances in Neural Information Processing Systems. pp. 785–792 (2002)
24. Hudson, D.: Interval estimation from the likelihood function. Journal of the Royal Statistical Society. Series B (Methodological) pp. 256–262 (1971)
25. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artificial Intelligence 172, 1897–1916 (2008)

26. King, R., Feng, C., Sutherland, A.: Statlog: Comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence 9(3), 289–333 (1995)
27. Lorena, A., de Carvalho, A., Gama, J.: A review on the combination of binary classifiers in multiclass problems. Artificial Intelligence Review 30(1), 19–37 (2008)
28. Marden, J.: Analyzing and modeling rank data, vol. 64. Chapman & Hall/CRC (1996)
29. Meila, M., Chen, H.: Dirichlet process mixtures of generalized mallows models. In: Proceedings of the 26th Conference in Uncertainty in Artificial Intelligence. pp. 358–367 (2010)
30. Troffaes, M.: Decision making under uncertainty using imprecise probabilities. Int. J. of Approximate Reasoning 45, 17–29 (2007)
31. Walley, P., Moral, S.: Upper probabilities based only on the likelihood functions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61, 831–847 (1999)
32. Weskamp, N., Hüllermeier, E., Kuhn, D., Klebe, G.: Multiple graph alignment for the structural analysis of protein active sites. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 4(2), 310–320 (2007)
33. Zaffalon, M.: Exact credal treatment of missing data. Journal of Statistical Planning and Inference 105(1), 105–122 (2002)