Soft Comput (2016) 20:939–950
DOI 10.1007/s00500-014-1553-9

METHODOLOGIES AND APPLICATION

Modelling and predicting partial orders from pairwise belief functions

Marie-Hélène Masson · Sébastien Destercke · Thierry Denoeux

Published online: 14 December 2014

Abstract In this paper, we introduce a generic way to represent and manipulate pairwise information about partial orders (representing rankings, preferences, ...) with belief functions. We provide generic and practical tools to make inferences from this pairwise information and illustrate their use on the machine learning problems that are label ranking and multi-label prediction. Our approach differs from most other quantitative approaches handling complete or partial orders, in the sense that partial orders are here considered as primary objects and not as incomplete specifications of ideal but unknown complete orders.

Keywords Dempster–Shafer theory · Belief functions · Paired comparisons · Partial orders · Label ranking · Multilabel classification

1 Introduction

The need to quantitatively model order structures and make inference about them is present in many fields: rank manipulation in statistics (Marden 1995), preference modelling in multi-criteria decision making (Grabisch and Labreuche 2008), preference learning (Fürnkranz and Hüllermeier 2010), decision theory, etc.

Orders being complex structures, information about them is often incomplete or uncertain. However, while the need to consider partial observations of complete orders in the previously mentioned fields have been quickly acknowledged, quantitative methods still focus mainly on representing and inferring complete orderings (this is in contrast with more qualitative representations, such as CP-nets, Boutilier et al. 2004, that are tailored to model partial orders).

It is only recently that the task of inferring partial orders has gained some interest in fields such as learning to rank (Cheng et al. 2010, 2012) or learning multi-criteria aggregation functions (Labreuche 2010) (note that this task of inferring partial orders is quite different from trying to choose a unique representation from partial informations, Greco et al. 2011). Even in these cases, partial orders are seen as incomplete but reliable inferences concerning ideals underlying complete orders, about which we have insufficient information.

This paper takes a quite different perspective, as partial orders are here considered as the primary objects, meaning that linear orders are merely a special case of the presented framework. More precisely, this work proposes to use belief functions bearing on the relation between pairs of objects and to infer information about partial orders from them. It provides generic as well as pragmatic tools to perform inferences. Belief functions indeed provide interesting uncertainty models to model partial information about orders: they allow to mix imprecise observations and belief degrees in a single setting, thus formally putting sets and probabilities under a common umbrella. One of the main contributions of this paper is to provide a framework where the incomparability relation of partial orders is explicitly modelled (in our case as a specific focal element), a feature that, as far as

Communicated by V. Loia.

M.-H. Masson
Université de Picardie Jules Verne, Amiens, France

M.-H. Masson (✉) · S. Destercke · T. Denoeux
UMR CNRS 7253 Heudiasyc, Université de Technologie de Compiègne, CS 60319, 60203 Compiègne Cedex, France
e-mail: mmasson@hds.utc.fr

S. Destercke
e-mail: sebastien.destercke@hds.utc.fr

T. Denoeux
e-mail: thierry.denoeux@hds.utc.fr

we know, is not present in works dealing with quantitative models of orders.

There are only few other works that deal with belief functions defined over partial orders. Among them, the work of Tritchler and Lockwood (1991) is probably the oldest one, but it remains very theoretical (not providing practical inference tools) and does not explicitly model incomparability. Utkin (2009) also proposes belief functions to work with partial orders, but considers frequencies of pairwise comparisons between groups of objects (i.e., imprecise observations) and no incomparability, while we consider pairwise comparisons between single objects and models incomparability, without necessarily referring to a frequentist interpretation.

The main framework we use is presented in Sect. 3, and is then illustrated in Sect. 4 on two problems of machine learning: label ranking and multilabel classification. Section 2 recalls the necessary elements of belief function theory.

2 Belief functions

The Theory of Belief Functions (also referred to as Dempster–Shafer or Evidence Theory) has been proposed by Shafer (1976) as a general model of uncertainties. By mixing probabilistic and set-valued representations, it allows to represent degrees of belief and incomplete information in a unified framework, which makes it adequate to model uncertainty about orders or preferences between objects. Indeed, preferences are most often partially observed and may be subject to various uncertainties (e.g., a decision maker can be quite uncertain about her/his preferences, or the preferences between different agents may be based on quite different amount of data).

Let us consider an uncertain variable ω taking values in a finite and unordered set Ω (in our case, this will be the set of asymmetric relations bearing on a set $\Lambda = \{\lambda_1, \ldots, \lambda_c\}$ of c objects) called the frame of discernment. Within belief function theory, the belief or the information regarding the actual value taken by ω is represented by a mass function Shafer (1976) and Smets and Kennes (1994) defined as a function m^Ω from 2^Ω to [0, 1], verifying

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1, \tag{1}$$

and

$$m^\Omega(\emptyset) = 0. \tag{2}$$

The notation m^Ω will be simplified to m when there is no ambiguity about the frame of discernment. The sets A of Ω such that $m(A) > 0$ are called *focal sets* of m. Each focal set A represents a set of possible values for ω, and the quantity $m(A)$ can be interpreted as a fraction of a unit mass of belief, which is allocated to A on the basis of a given evidential corpus. Complete ignorance corresponds to $m(\Omega) = 1$ (vacuous mass), and perfect knowledge of the value of ω is represented by the certain mass assigning the whole mass of belief to a unique singleton of Ω. A mass function is said to be *logical* if it has only one focal set. A mass function is *simple* if it has at most two focal sets, including Ω.

Any mass function can be equivalently represented by a belief function bel, a plausibility function pl and a commonality function q defined, respectively, by

$$\mathrm{bel}(A) \triangleq \sum_{B \subseteq A} m(B), \tag{3}$$

$$\mathrm{pl}(A) \triangleq \sum_{B \cap A \neq \emptyset} m(B), \tag{4}$$

$$\mathrm{q}(A) \triangleq \sum_{B \supseteq A} m(B), \tag{5}$$

for all $A \subseteq \Omega$. The belief function quantifies how much event A is implied by the information, as it sums up masses of sets included in A and whose mass is necessarily allocated to A. The plausibility function quantifies how much event A is consistent with our information, as it sums masses that do not contradict A, i.e., whose intersection with A is non-empty. The commonality is harder to interpret, but possesses mathematical properties that we will exploit in this paper. The first of them is that for any singleton $\omega \in \Omega$, we have $\mathrm{pl}(\{\omega\}) = \mathrm{q}(\{\omega\})$.

Two mass functions m_1 et m_2 on Ω representing two distinct pieces of evidence may be combined by Dempster's rule of combination. The resulting mass function $m_\oplus = m_1 \oplus m_2$ is given by

$$m_\oplus(A) \triangleq \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C), \tag{6}$$

with K, called the degree of conflict, defined as the mass $m(\emptyset)$ given to the empty set, i.e.,

$$K \triangleq \sum_{B \cap C = \emptyset} m_1(B) m_2(C). \tag{7}$$

Dempster's rule of combination may be equivalently expressed using commonalities as

$$q_\oplus(A) = \frac{1}{1-K} \mathrm{q}_1(A) \mathrm{q}_2(A) \quad \forall A \subseteq \Omega, \tag{8}$$

where $\mathrm{q}_1(A)$ and $\mathrm{q}_2(A)$ are, respectively, the commonalities associated to m_1 and m_2. This provides a very easy means to express the result of Dempster's rule on events $A \subseteq \Omega$, and we will make heavy use of it in the next sections.

3 Partial order prediction in the framework of belief functions using a pairwise approach

3.1 Problem statement, notations

Let $\Lambda = \{\lambda_1, \ldots, \lambda_c\}$ denote the set of c possible objects (labels, alternatives, etc.). A binary relation $R \subseteq \Lambda \times \Lambda$ is asymmetric if $(\lambda_i, \lambda_j) \in R \Rightarrow (\lambda_j, \lambda_i) \notin R$, and we will denote by $\mathcal{A}$ the set of asymmetric relations defined over Λ.

A strict partial order over Λ is a binary relation that is transitive[1] and asymmetric, and we will denote by $\mathcal{R}$ the set of such partial orders over Λ. In this latter case $(\lambda_i, \lambda_j) \in R$, also written $\lambda_i \succ_R \lambda_j$, indicates that λ_i is strictly higher than (preferred to) λ_j. If neither $\lambda_i \succ_R \lambda_j$ nor $\lambda_j \succ_R \lambda_i$ holds, λ_i and λ_j are said to be incomparable and we will denote it by $\lambda_i \prec\succ_R \lambda_j$.

A partial order R is a total (or linear) order if it satisfies the additional completeness property stating that, for every $\lambda_i, \lambda_j \in \Lambda$, either $\lambda_i \succ_R \lambda_j$ or $\lambda_j \succ_R \lambda_i$ must hold. As explained in the introduction, one of the main contributions of this paper is to introduce various means to perform inferences on the space $\mathcal{R}$ or on one of its subspace, given that our initial information comes in the form of belief functions bearing on the $c(c-1)/2$ pairs λ_i, λ_j, $i < j$ of labels.

Considering such pairwise decomposition is common when dealing with orders, as the space we are working on quickly becomes huge as c increases. Indeed, with $c = 8$, there are already 431,723,379 possible partial order relations (source *On-line Encyclopedia of Integer Sequences*, http://oeis.org/A001035) and $8! = 40,320$ linear orders. Directly working on $\mathcal{R}$ would be computationally prohibitive and hence the interest of decomposing the problem into a set of simpler ones.

3.2 Combination of pairwise belief functions

For each pair $\{\lambda_i, \lambda_j\}$, $i < j$ we consider that the information about the relation between λ_i and λ_j provided by some source (e.g., a classifier as in Sect. 4, a decision maker, a recommendation system) is expressed using a mass function quantifying the uncertainty related to this relation and denoted by $m^{\Theta_{ij}}$. This mass function has the following general form:

$$\begin{cases} m^{\Theta_{ij}}(\lambda_i \succ \lambda_j) = \alpha_{ij}, \\ m^{\Theta_{ij}}(\lambda_j \succ \lambda_i) = \beta_{ij}, \\ m^{\Theta_{ij}}(\lambda_i \prec\succ \lambda_j) = \gamma_{ij}, \\ m^{\Theta_{ij}}(\mathcal{A}) = 1 - \alpha_{ij} - \beta_{ij} - \gamma_{ij}, \end{cases} \tag{9}$$

with $\lambda_i \succ \lambda_j$, $\lambda_j \succ \lambda_i$, and $\lambda_i \prec\succ \lambda_j$ being short notations for the events $\{R \in \mathcal{A} : (\lambda_i, \lambda_j) \in R\}$, $\{R \in \mathcal{A} : (\lambda_j, \lambda_i) \in R\}$, and $\{R \in \mathcal{A} : (\lambda_i, \lambda_j) \notin R \text{ and } (\lambda_j, \lambda_i) \notin R\}$, respectively. Note that the above mass may give a positive weight to incomparability, which implicitly means that such incomparability can be observed. While this is a reasonable assumption in some cases, such as when decision makers are unable to compare two alternatives, or in learning problems where observations are partial orders (e.g., multilabel prediction, see Sect. 4), in others (e.g., label ranking) assuming incomparability can be observed may turn out to be unreasonable, in which case one can simply impose $\gamma_{ij} = 0$ in (9).

When the observation is the preference $\lambda_i \succ \lambda_j$ with some reliability α_{ij}, then this can be modelled by setting $\gamma_{ij} = 0$ and $\beta_{ij} = 0$ in Expression (9). Similarly, observing the preference $\lambda_j \succ \lambda_i$ with some reliability β_{ij} can be modelled by setting $\gamma_{ij} = 0$ and $\alpha_{ij} = 0$.

The combination of the $c(c-1)/2$ pairwise mass functions defined by (9) may be seen as an information fusion problem and Dempster's rule of combination may be used to this end. Applying Dempster's rule yields

$$m^{\mathcal{A}} = m^{\Theta_{12}} \oplus m^{\Theta_{13}} \oplus \cdots \oplus m^{\Theta_{(n-1)n}}, \tag{10}$$

This combination can be computed equivalently using the commonalities by

$$\mathrm{q}^{\mathcal{A}} = \prod_{i<j} \mathrm{q}^{\Theta_{ij}}. \tag{11}$$

Since the focal elements of two distinct masses $m^{\Theta_{ij}}$ and $m^{\Theta_{k\ell}}$ will contain information about different pairs of labels $\{\lambda_i, \lambda_j\}$, $\{\lambda_k, \lambda_\ell\}$, the intersections between the focal elements of $m^{\Theta_{ij}}$ and of $m^{\Theta_{k\ell}}$ will be non-empty, and the results of these intersections will be asymmetric relations. In particular, this also means that in our case the value of K in (6) will be null.

Example 1 Let $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$. Let us consider the following focal elements, respectively, taken from $m^{\Theta_{12}}$, $m^{\Theta_{13}}$ and $m^{\Theta_{23}}$

- $\lambda_1 \succ \lambda_2 := \{\{(\lambda_1, \lambda_2), (\lambda_2, \lambda_3), (\lambda_1, \lambda_3)\}, \{(\lambda_1, \lambda_2), (\lambda_2, \lambda_3), (\lambda_3, \lambda_1)\}, \{(\lambda_1, \lambda_2), (\lambda_3, \lambda_2), (\lambda_1, \lambda_3)\}, \{(\lambda_1, \lambda_2), (\lambda_2, \lambda_3), (\lambda_1, \lambda_3)\}, \{(\lambda_1, \lambda_2), (\lambda_3, \lambda_2)\}, \{(\lambda_1, \lambda_2), (\lambda_2, \lambda_3)\}, \{(\lambda_1, \lambda_2), (\lambda_1, \lambda_3)\}, \{(\lambda_1, \lambda_2), (\lambda_3, \lambda_1)\}, \{(\lambda_1, \lambda_2)\}\}$
- $\lambda_1 \prec\succ \lambda_3 := \{\{(\lambda_1, \lambda_2), (\lambda_2, \lambda_3)\}, \{(\lambda_1, \lambda_2), (\lambda_3, \lambda_2)\}, \{(\lambda_2, \lambda_1), (\lambda_2, \lambda_3)\}, \{(\lambda_2, \lambda_1), (\lambda_3, \lambda_2)\}, \{(\lambda_1, \lambda_2)\}, \{(\lambda_2, \lambda_1)\}, \{(\lambda_3, \lambda_2)\}, \{(\lambda_2, \lambda_3)\}\}$
- $\mathcal{A}$

The focal element ϕ resulting from the intersection of $\lambda_1 \succ \lambda_2$, $\lambda_1 \prec\succ \lambda_3$ and $\mathcal{A}$ is the set composed of the three following partial asymmetric relations:

$$\phi = \{\{(\lambda_1, \lambda_2), (\lambda_2, \lambda_3)\}, \{(\lambda_1, \lambda_2), (\lambda_3, \lambda_2)\}, \{(\lambda_1, \lambda_2)\}\}.$$

[1] $(\lambda_i, \lambda_j), (\lambda_j, \lambda_k) \in R \Rightarrow (\lambda_i, \lambda_k) \in R$.

The mass of ϕ resulting from the combination is

$$m^{\mathcal{A}}(\phi) = m^{\Theta_{12}}(\lambda_1 \succ \lambda_2) m^{\Theta_{13}}(\lambda_1 \prec\succ \lambda_3) m^{\Theta_{23}}(\mathcal{A}). \quad (12)$$

3.3 Inferences on partial orders

The combination (10) allocates masses to various sets of asymmetric relations defined on $\Lambda \times \Lambda$. However, the focus of this paper is on partial orders $\mathcal{R}$ (a subset of $\mathcal{A}$), and we will now explain how we can go from a model on the space $\mathcal{A}$ to inferences on $\mathcal{R}$. From (10), finding a partial order belonging to some specific set $\mathcal{S} \subseteq \mathcal{R}$ can be done by finding the element within $\mathcal{S}$ that has the maximum plausibility, that is by finding

$$\hat{\omega} = \underset{\omega \in \mathcal{S}}{\text{argmax}}\ \text{pl}^{\mathcal{A}}(\{\omega\}). \quad (13)$$

Using the maximum of plausibility on singletons has been recently proposed as an efficient decision tool in many problems involving belief functions (El Zoghby et al. 2013; Denœux and Masson 2012). It also corresponds to a maximax type of decision (thus adopting an "optimistic" criterion, Troffaes 2007) or to the decision that would be taken using the plausibility transform (Cobb and Shenoy 2006).

Rather than directly computing (13), we could first condition on the space $\mathcal{S}$ and then find the element of the conditioned belief function with the highest plausibility. Actually, the decision given by (13) is also consistent with such an approach. Indeed, recall that conditioning on a set B with Dempster's conditioning gives the conditioned plausibility measure $\text{pl}(A|B) = {}^{\text{pl}(A\cap B)}/_{\text{pl}(B)}$, hence for any $\omega, \omega' \in B$ if $\text{pl}(\omega) \leq \text{pl}(\omega')$, then ${}^{\text{pl}(\omega)}/_{\text{pl}(B)} \leq {}^{\text{pl}(\omega')}/_{\text{pl}(B)}$, meaning that the ordering on singleton plausibilities before or after conditioning remains unchanged.

Using the fact that plausibilities coincide with commonalities on singletons, plausibilities of singletons can be easily expressed as a product of plausibilities of combined belief functions using Eq. (11). Let us consider an element R of $\mathcal{R}$. As R is a singleton, we have that

$$\begin{cases} \text{q}^{\Theta_{ij}}(\{R\}) = 1 - \beta_{ij} - \gamma_{ij} & \text{if } \lambda_i \succ_R \lambda_j, \\ \text{q}^{\Theta_{ij}}(\{R\}) = 1 - \alpha_{ij} - \gamma_{ij} & \text{if } \lambda_j \succ_R \lambda_i, \\ \text{q}^{\Theta_{ij}}(\{R\}) = 1 - \alpha_{ij} - \beta_{ij} & \text{if } \lambda_j \prec\succ_R \lambda_i. \end{cases} \quad (14)$$

Using (14) and (8), the commonality/plausibility of R can be written as

$$\text{q}^{\mathcal{A}}(\{R\}) = \text{pl}^{\mathcal{A}}(\{R\}) \propto \prod_{\lambda_i \succ_R \lambda_j} (1 - \beta_{ij} - \gamma_{ij}) \times \prod_{\lambda_l \succ_R \lambda_k} (1 - \alpha_{kl} - \gamma_{kl}) \prod_{\lambda_m \prec\succ_R \lambda_n} (1 - \alpha_{mn} - \beta_{mn}). \quad (15)$$

Given the size of the space $\mathcal{R}$ (or one of its subset of interest), finding the most plausible order relation obviously cannot be done by enumeration. This is why we propose a generic approach to get this most plausible relation, this approach consisting of reformulating the problem as a binary integer linear programming problem to which can then be applied state-of-art techniques issued from optimization. First, we introduce the binary variables r_{ij}, $i, j = 1, \ldots c$ defined by

$$\begin{cases} \lambda_i \succ_R \lambda_j \iff r_{ij} = 1 \text{ and } r_{ji} = 0, \\ \lambda_i \prec\succ_R \lambda_j \iff r_{ij} = 0 \text{ and } r_{ji} = 0. \end{cases}$$

Let us now define, for each $i < j$:

$$\begin{cases} X_{ij}^{(1)} = r_{ij}(1 - r_{ji}), \\ X_{ij}^{(2)} = r_{ji}(1 - r_{ij}), \\ X_{ij}^{(3)} = (1 - r_{ij})(1 - r_{ji}). \end{cases}$$

Expression (15) can be rewritten using these new binary variables as

$$\text{pl}^{\mathcal{A}}(\{R\}) \propto \prod_{i<j} (1-\beta_{ij}-\gamma_{ij})^{X_{ij}^{(1)}} (1-\alpha_{ij}-\gamma_{ij})^{X_{ij}^{(2)}} (1-\alpha_{ij}-\beta_{ij})^{X_{ij}^{(3)}}. \quad (16)$$

Maximizing expression (16) is equivalent to maximizing its logarithm so that the most plausible relation $\hat{\omega} \in \mathcal{A}$ may be found as the solution of the following binary integer programming problem:

$$\begin{aligned} \max_{X_{ij}^{(k)} \in \{0,1\}} & \sum_{i<j} X_{ij}^{(1)} \ln(1 - \beta_{ij} - \gamma_{ij}) \\ & + \sum_{i<j} X_{ij}^{(2)} \ln(1 - \alpha_{ij} - \gamma_{ij}) \\ & + \sum_{i<j} X_{ij}^{(3)} \ln(1 - \alpha_{ij} - \beta_{ij}). \end{aligned} \quad (17)$$

subject to

$$\sum_{k=1}^{3} X_{ij}^{(k)} = 1 \quad \forall i < j, \quad (18)$$

this constraint ensuring that only one alternative between $\lambda_i \succ \lambda_j$, $\lambda_j \succ \lambda_i$ and $\lambda_i \prec\succ \lambda_j$ will be chosen. Depending on the nature of the relation we want to retrieve (or condition on), additional constraints may be imposed, as we shall now detail for partial orders, linear orders and bipartite rankings.

Partial orders the property of transitivity satisfied by partial orders can be encoded by the following constraint:

$$r_{ij} + r_{jk} - 1 < r_{ik} \quad \forall i, j, k, \quad (19)$$

that has to be added to (18) to ensure that the solution of (17) will be a partial order. It can be expressed using the variables of the problem since the following equations hold:

$$r_{ij} = \begin{cases} 1 - X_{ij}^{(2)} - X_{ij}^{(3)} & \text{if } i < j, \\ 1 - X_{ij}^{(1)} - X_{ij}^{(3)} & \text{if } i > j. \end{cases} \tag{20}$$

Linear orders a linear order is a partial order without incomparability. Basically, there are two ways within our framework to obtain linear orders: either set $\gamma_{ij} = 0$ for all $i \leq j$ in (9) and solve (17) under constraints (18)–(19), so that partial orders with incomparability cannot have the highest plausibility, or simply add the constraint

$$X_{ij}^{(3)} = 0 \quad \forall i < j \tag{21}$$

stating that λ_i, λ_j cannot be incomparable.

Bipartite rankings a (partial) bipartite ranking consists in dividing the objects or labels in two subsets: the preferred ones and the non-preferred ones. In practice, this means that there cannot be three objects $\lambda_i, \lambda_j, \lambda_k$ such that $\lambda_i \succ \lambda_j \succ \lambda_k$, as there are only two subsets. Such kind of partial orders are at work, e.g., in multilabel classification problems. This can be encoded by adding the constraints

$$r_{ij} + r_{jk} \leq 1 \quad \forall i, j, k. \tag{22}$$

to constraint (18). Multipartite ranking, where K ordered subsets $L_1, \ldots, L_k$ of objects must be constructed, can be modelled by extending constraints (22) to more than triplets of objects and by adding transitivity constraints.

Partial orders, linear orders and multipartite rankings are only some examples illustrating the flexibility of our framework, since many constraints are very easy to formulate within it. For instance, it is straightforward to impose some (known) relation to hold, simply by constraining the variables $X_{ij}^{(k)}$ to adequate values. It would also be interesting to investigate if some specific families of partial orders, such as interval orders or semi-orders, can be easily expressed through constraints.

4 Applications

We now illustrate how our approach can be applied to some well-known and difficult machine learning problems, namely label ranking and multilabel classification. These kinds of problems find applications in various domains like music or text categorization, bioinformatics, semantic classification of images, recommendation systems (Boutell et al. 2004; Elisseeff and Weston 2001; Li and Ogihara 2006; Ueda and Saito 2002, ...). In particular, we explore how partial rankings can be predicted for the label ranking problem using our approach and the interest of explicitly modelling the incomparability in the multilabel problem.

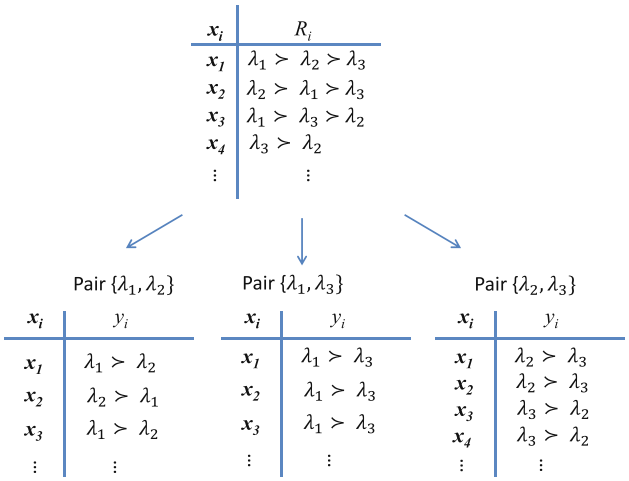

Fig. 1 Pairwise decomposition in case of label ranking

In traditional (single-label) classification, it is assumed that each observation $\mathbf{x}$ of an input space $\mathcal{X}$ is associated with a single label λ from a finite possible set $\Lambda = \{\lambda_1, \ldots, \lambda_c\}$. The task is then to learn the mapping or classifier from $\mathcal{X}$ to Λ using a training set of n observations $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i$ in $\mathcal{X}$ and $y_i \in \Lambda$. Label ranking and multilabel settings differ from the assumptions of the traditional one in the sense that it is assumed that with each observation is associated a linear order (possibly imprecisely observed) in the case of label ranking and a set of relevant labels defining a bipartite ranking in the case of multilabel prediction.

4.1 Label ranking

In the label ranking setting Vembu and Gärtner (2011), each instance $\mathbf{x}_i$ is associated with a linear order R_i over Λ, possibly partially observed. Rather than training one classifier over all possible orderings (whose number is $c!$), a common strategy is to decompose the problem into pairwise preferences, and to train $c(c-1)/2$ classifiers, that is one for each pair of labels (Hüllermeier et al. 2008). Other approaches propose to estimate the mapping between instances and complete or partial rankings by learning a utility function for each label (e.g., by constraint classification or log-linear models), from which a ranking can then be deduced (Dekel et al. 2003; Har-Peled et al. 2003).

The pairwise decomposition applied to an illustrative data set is pictured by Fig. 1. From this figure, one can see that this decomposition tackles the problem of missing data in a straightforward way, as missing pairwise preferences will simply mean less data in the corresponding split.

In the usual label ranking setting, the goal is to predict a linear order as close as possible to the observed one (note that, in the case of label ranking, the notion of closeness can be modelled by various loss functions, Hüllermeier et al. 2008).

However, some authors have recently discussed the interest of producing not linear orders, but partial orders as predictions (Cheng et al. 2010). The idea behind such predictions is close to the one of the reject option (Chow 1970) in traditional classifiers or to credal classification (Zaffalon 2002), that is to abstain to make precise predictions when the available information is insufficient.

As our framework is tailored to infer partial orders, we first apply it in this context of label ranking with partial abstention and compare it to the approach of Cheng et al. (2010), first considered by Rademaker and De Baets (2010) in the general setting of partial order aggregation. The method of Cheng et al. is based on the principle of classifier ensembles. Instead of training a single model, the idea is to train k binary probabilistic classifiers by creating k bootstrap samples from the original data. In this way, k estimates $\hat{p}_{ij}^{\ell}, \ell = 1, \ldots, k$ of the probability $\mathbb{P}(\lambda_i \succ \lambda_j)$ are thus available. For each pair of labels $\{\lambda_i, \lambda_j\}$, a preference degree $P(\lambda_i, \lambda_j)$ is defined as the fraction of classifiers for which λ_i is preferred λ_i to λ_j:

$$P(\lambda_i, \lambda_j) = \frac{1}{k} |\{\ell : \hat{p}_{ij}^{\ell} > 0.5\}|. \tag{23}$$

Then, a binary relation is derived by thresholding the preference degrees:

$$R_\alpha = \{(\lambda_i, \lambda_j) : P(\lambda_i, \lambda_j) \geq \alpha\}, \tag{24}$$

where the threshold α is such that $\alpha > \frac{\lfloor k/2 \rfloor}{k}$, where $\lfloor x \rfloor$ denotes the integer part of x. Note that, by construction, the number of possibly different values in P, and consequently the number of possible thresholds, is directly related to the number of bootstrap replicates. The authors underline that, for a given threshold, especially if it is low, the relation obtained is not necessarily transitive and may even contain cycles. The transitivity may be easily enforced by computing the transitive closure, but guaranteeing the absence of cycle is more problematic. The authors propose a procedure to find the minimum value of α such that the transitive closure of the partial relation is a strict partial order relation. By varying the value of the threshold, different relations are obtained: the larger the α value, the less informative the corresponding relation (i.e., the more incomparabilities it contains).

To compare our approach with the one of Cheng et al. (2010), we follow the following procedure:

1. we build k bootstrap replicates of the learning set;
2. from each bootstrap sample, we train $c(c-1)/2$ classifiers;
3. for a fixed threshold τ, we compute m_{ij}^{Θ} as follows:

$$\begin{cases} m^{\Theta_{ij}}(\lambda_i \succ \lambda_j) = \frac{1}{k} |\{l : \hat{p}_{ij}^{l} > 0.5 + \tau\}|, \\ m^{\Theta_{ij}}(\lambda_j \succ \lambda_i) = \frac{1}{k} |\{l : \hat{p}_{ij}^{l} < 0.5 - \tau\}|, \\ m^{\Theta_{ij}}(\lambda_i \prec\succ \lambda_j) = \frac{1}{k} |\{l : 0.5 - \tau < \hat{p}_{ij}^{l} < 0.5 + \tau\}|, \end{cases} \tag{25}$$

where it is clear that the higher τ is, the more incomparabilities are favoured.

4. To take account of the different performances of each classifier, a discounting operation (Shafer 1976) is applied to each mass $m^{\Theta_{ij}}$. For a given mass, let $\epsilon \in [0, 1]$ be the average of the k values of mean squared error obtained on the training set; then $m^{\Theta_{ij}}$ is transformed into ${}^{\epsilon}m^{\Theta_{ij}}$ such that

$$\begin{cases} {}^{\epsilon}m^{\Theta_{ij}}(A) = (1 - \epsilon) m(A) \quad \forall A \neq \mathcal{A}, \\ {}^{\epsilon}m^{\Theta_{ij}}(\mathcal{A}) = (1 - \epsilon) m(\mathcal{A}) + \epsilon, \end{cases} \tag{26}$$

The coefficient ϵ is usually interpreted as the unreliability of the source of information (Smets 1993): when ϵ is equal to 1, m becomes the vacuous mass function, and when $\epsilon = 0$ the information is fully reliable and m remains unchanged.

5. The most plausible partial order is computed by solving problem (17) under constraints (18)–(19). By varying the value of τ ($\tau \in [0; 0.5[$), different partial orders are obtained: the larger the τ value, the less informative the corresponding relation.

Table 1 Data sets description for label ranking

Data set	# Features	# Labels	# Instances
Iris	4	3	150
Wine	13	3	178
Glass	9	6	214
Vehicle	18	4	846
Pendigits	26	10	2,199
Segment	18	7	2,310
Cold	24	4	2,465
Heat	24	6	2,465

We applied this strategy to eight data sets presented in Hüllermeier et al. (2008). The first six datasets are multiclass data sets from the UCI machine learning repository (Bache and Lichman 2013) that have been transformed into label ranking data by a procedure proposed in Hüllermeier et al. (2008). A naive Bayes classifier was trained on the entire data set. Then, for each instance, the labels were ordered according to the predicted class probabilities.[2] The last two datasets (cold and heat) are real-world data sets originating from the bioinformatics field and are described in Hüllermeier et al. (2008). All these data sets are described in Table 1.

To evaluate the performance of the methods, we use two measures proposed in Cheng et al. (2010). The first one, *correctness*, quantifies how the predicted (partial) ranking

[2] The data sets are available at http://www.uni-marburg.de/fb12/kebi/research/repository/labelrankingdata.

matches the observed ranking, whereas the second one is intended to measure the degree of *completeness* of the relation. A good method predicting partial orders should see its correctness increase as the completeness decrease, and there is usually a trade-off to find between these two criteria. They can be formally defined as follows: let $(\mathbf{x}_i, L_i)$, $i = 1, n$, denote the test set. L_i is the true linear order relation on $\Lambda \times \Lambda$ and let R_i denote the partial (or linear) order computed by the method. For each $\mathbf{x}_i$, a pair of labels $\{\lambda_k, \lambda_l\}$ is said to be concordant if

$$((\lambda_k, \lambda_l) \in L_i \text{ and } (\lambda_k, \lambda_l) \in R_i) \text{ or } ((\lambda_l, \lambda_k) \in L_i \text{ and } (\lambda_l, \lambda_k) \in R_i).$$

It is said to be discordant if

$$((\lambda_k, \lambda_l) \in L_i \text{ and } (\lambda_l, \lambda_k) \in R_i) \text{ or } ((\lambda_l, \lambda_k) \in L_i \text{ and } (\lambda_k, \lambda_l) \in R_i).$$

Let c_i and d_i denote the number of concordant and discordant pairs of alternatives for sample $(\mathbf{x}_i, L_i)$, respectively. The correctness measure for the test set is defined as

$$\text{Correctness} = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i - d_i}{c_i + d_i},$$

whereas the completeness is defined by

$$\text{Completeness} = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i + d_i}{n(n-1)/2}.$$

The performances of our method were compared to those of Cheng et al. (2010). For both methods, the base learner was a logistic regression. The number of bootstrap replicates was fixed to 11. The results reported in Fig. 2 are the mean values computed over five repetitions of a tenfold cross-validation procedure. As already explained, for some instances, Cheng's et al. method fails at finding an acyclic relation for low values of threshold, unlike our method which is able to find a solution in any situation. As an illustration, Table 2 gives the average rate of non-responses of Cheng's method, computed over five repetitions for different data sets with $\alpha = 0.5$. The instances for which no relation is found are usually complex cases. So, for the comparison to be fair, completeness and correctness are computed only over instances for which a relation is found by both methods.

Our results and the results of Cheng et al. are represented, respectively, by solid and dashed lines. From these experiments, the following conclusions may be drawn:

1. As expected, for both methods, partial abstention leads to improved correctness, that is the abstention is done on poorly reliable predictions. For some data sets, a significant gain in performance can be achieved without losing too much completeness.
2. The performances of the two methods are very similar for large values of completeness, but it can be seen that our approach (1) usually provides better correctness results when completeness decreases and (2) is able to span a wider range of completeness values, as it can go from linear to vacuous order.

Another, perhaps surprising observation is that providing correct predictions seems much more difficult on the genuine label ranking data sets (on which our method clearly performs better when allowing for partial orders) than on the synthetic label ranking data sets. This suggests that the synthetic data sets are much more regular than genuine label ranking problems, in which case it appears important to consider the two kind of sets when assessing label ranking methods.

Note that we have checked that always giving a response does not decrease dramatically the performance of our method. For example, looking for a complete relation, the correctness degree for the iris data set is equal to 0.88 when it is computed using the entire data set instead of 0.9, and to 0.86 instead of 0.88 for the glass data set.

4.2 Multilabel classification

In the multilabel setting (Tsoumakas et al. 2010), each instance can be associated simultaneously with one or several labels. This association implicitly defines a bipartition of the labels into relevant and irrelevant labels. The task we consider in this section is thus to learn a mapping from $\mathcal{X}$ to the powerset of Λ from the training data $(\mathbf{x}_i, Y_i)$, $i = 1, \ldots, n$ with $\mathbf{x}_i$ in $\mathcal{X}$ and $Y_i \subseteq \Lambda$.

In recent years, many approaches have been suggested to solve this problem. A review of these methods can be found in Tsoumakas and Katakis (2007) and Tsoumakas et al. (2010). Also, Madjarov et al. (2012) have proposed an extensive experimental comparison of the methods. In their paper, the performances of 12 methods are studied according to different evaluation measures/loss functions. The methods are divided into three categories: algorithm adaptation, problem transformation and ensemble methods. Algorithm adaptation methods extend existing machine learning algorithms to the problem of multi-label classification. Problem transformation methods are multi-label learning methods that transform the multi-label problem into one or more single-label problems, like the pairwise approach used in this paper. Ensemble methods use problems transformation methods or algorithms adaptation as base classifiers. A summary of all these methods is given in Table 3.

Classically, the transformation of the multilabel problem into an ordering problem somehow assumes that observations Y_i are incomplete observations of complete rankings (Fürnkranz et al. 2008). The task is then to learn how to predict complete rankings out of these observations, and

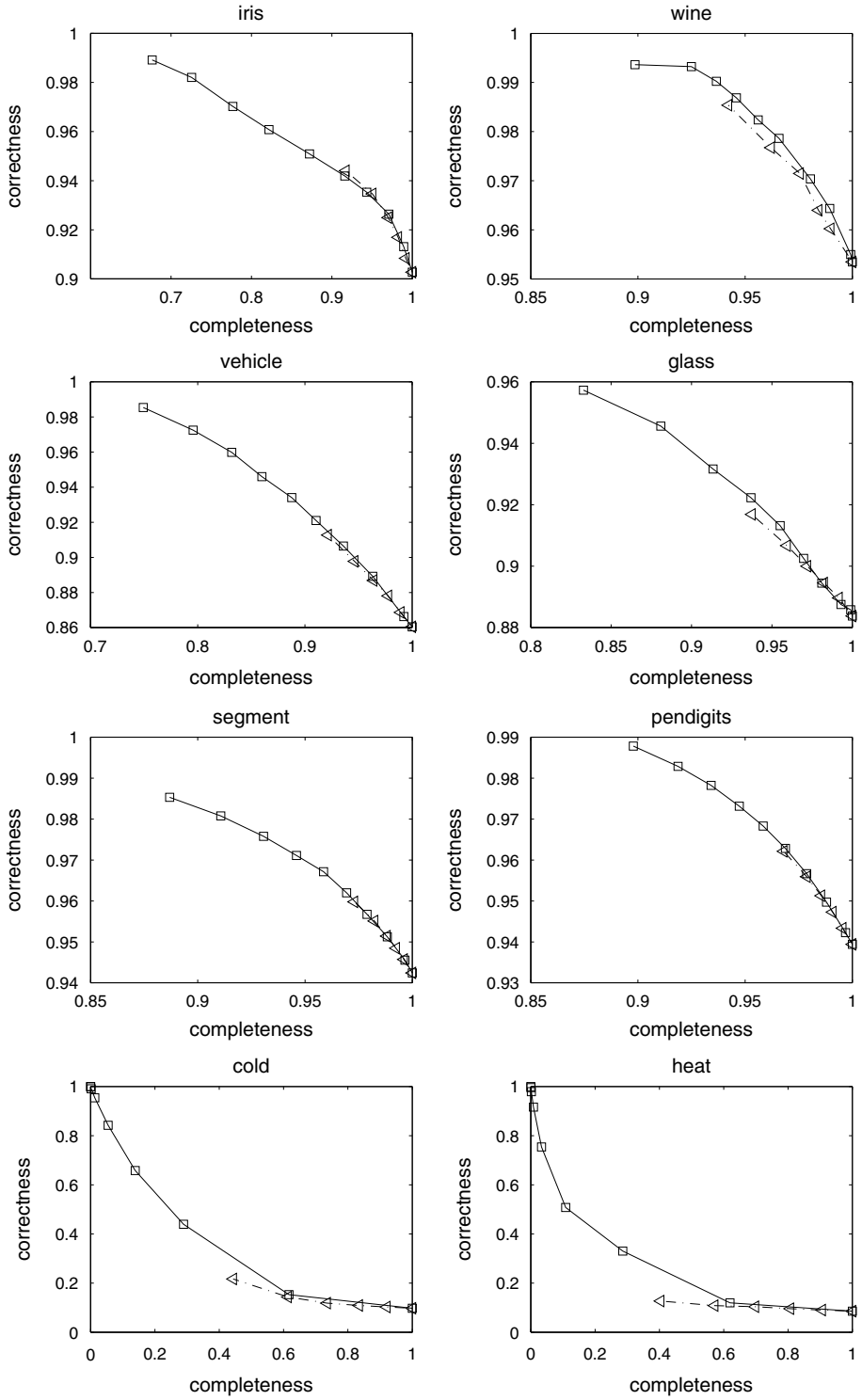

Fig. 2 Correctness versus completeness for different data sets (*dashed line*, *triangles*: Cheng et al. method; *solid line*, *squares*: our method)

a multilabel prediction is then obtained by "cutting off" the order between relevant and irrelevant labels.

Our approach is different, and as far as we know has not been proposed before in a multilabel setting: we try to predict not a complete order, but directly a partial order corresponding to a bipartite ranking. In particular, this means that we include incomparabilities as observations. More precisely, in the training set labels incomparability between λ_j, λ_k will be observed in Y_i if either $\lambda_j, \lambda_k \in Y_i$ or $\lambda_j, \lambda_k \notin Y_i$; otherwise $\lambda_j \succ \lambda_k$ when $\lambda_j \in Y_i, \lambda_k \notin Y_i$. A noticeable difference with viewing multilabel observations as incompletely observed linear ordering is that in our case, if all multilabel

Table 2 Non-response rate for Cheng's et al. method ($\alpha = 0.5$)

Data set	Rate (%)
Iris	5.47
Wine	0.2
Glass	5.51
Vehicle	1.61
Authorship	0.2
Segment	6.55

observations are precise, all pairwise training data sets contain as many data as the original data set, that is there are no "missing" preferences (Fig. 3).

We compare the results given in Madjarov et al. (2012) with the results obtained with our method using three usual data sets: emotions, scene and yeast, which are described in Table 4. The performances of the methods are evaluated using four popular evaluation criteria for multi-label classification: the ranking loss, the Hamming loss, the F_1-measure and the accuracy. Let $(\mathbf{x}_i, Y_i)$, $i = 1, \ldots, n$, $Y_i \subseteq \Lambda$, denote the set of test samples. Let $\widehat{Y}_i$ denote the set of predicted labels for instance $\mathbf{x}_i$. The evaluation criteria are defined as follows:

The Hamming loss evaluates how many times a label is misclassified, i.e., a label not belonging to the instance is predicted or a label belonging to the instance is not predicted. The measure has to be minimized (performance is perfect when Hammingloss $= 0$) and is defined by

$$\text{Hamming loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \Delta \widehat{Y}_i|}{c}, \tag{27}$$

where Δ stands for the symmetric difference between two sets and c for the number of labels.

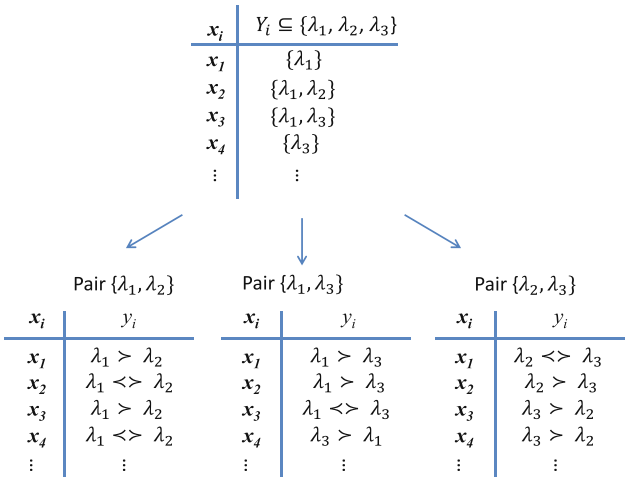

Fig. 3 Pairwise decomposition in case of multilabel classification

The accuracy is the average over the test samples of the Jaccard coefficient between the sets Y_i and $\widehat{Y}_i$; it has to be maximized:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap \widehat{Y}_i|}{|Y_i \cup \widehat{Y}_i|} \tag{28}$$

F_1 score is the harmonic mean between two other evaluation measures, namely precision and recall, and is computed as follows:

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2|Y_i \cap \widehat{Y}_i|}{|Y_i| + |\widehat{Y}_i|} \tag{29}$$

The best value for F_1 is 1 and the worst is 0. The ranking loss, which has to be minimized, evaluates the average fraction of pairs of labels that are reversely ordered:

Table 3 Methods for multilabel classification

Method	References	Type
Binary relevance	Tsoumakas and Katakis (2007)	Pb transformation (one vs all)
Classifier chaining	Read et al. (2011)	Pb transformation (one vs all)
HOMER	Tsoumakas et al. (2008)	Pb transformation (label powerset)
Calibrated label ranking	Fürnkranz et al. (2008)	Pb transformation (pairwise)
QWML	Loza Mencía et al. (2010)	Pb transformation (pairwise)
Multilabel C4.5 (ML C4.5)	Clare and King (2001)	Algorithm adaptation
Multilabel kNN	Zhang and Zhou (2007)	Algorithm adaptation
Predictive clustering trees (PCT)	Blockeel et al. (1998)	Algorithm adaptation
ECC	Read et al. (2011)	Ensemble methods
Random forest ML-C4.5	Madjarov et al. (2012)	Ensemble methods
Random forest of PCT	Kocev et al. (2007)	Ensemble methods
Rakel	Tsoumakas and Vlahavas (2007)	Ensemble methods

Table 4 Data sets description for multilabel classification

Data set	# Features	# Labels	# Tr. instances	# T. instances	Cardinality
Scene	294	6	1,211	1,159	1.07
Emotions	72	6	391	202	1.87
Yeast	103	14	1,500	917	4.24

Table 5 Results for the Scene dataset

	Worst	Best	BFPC	Rank
F1 (max)	0.395	0.771	0.6947	8
HL (min)	0.141	0.077	0.1048	10
RL (min)	0.174	0.064	0.1127	10
ACC (max)	0.388	0.735	0.6813	8

Table 6 Results for the Yeast dataset

	Worst	Best	BFPC	Rank
F1 (max)	0.578	0.687	0.6374	8
HL (min)	0.234	0.190	0.2054	9
RL (min)	0.296	0.163	0.0846	1
ACC (max)	0.440	0.559	0.5357	3

$$\text{Ranking loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{|D_i|}{|Y_i||\widehat{Y}_i|}, \tag{30}$$

where $D_i = \{(\lambda_i, \lambda_j)|(\lambda_i, \lambda_j) \in Y_i \times \overline{Y_i}$ and $(\lambda_i, \lambda_j) \in \overline{\widehat{Y}_i} \times \widehat{Y}_i\}$. Note that this D_i is formally equivalent to the number of discordant pairs used for evaluating label ranking in Sect. 4.1, but is here used in a different way (in particular, there is no need to split between correctness and completeness, as all predictions are complete).

As binary classifiers, we used the evidential kNN (Denœux 1995) method with three classes ($\lambda_i \succ \lambda_j, \lambda_j \succ \lambda_i, \lambda_i \prec\succ \lambda_j$), which directly provides a mass function in a form of (25).[3] The optimum number of nearest neighbours (the same for all classifiers) is determined using five repetitions of tenfold cross-validation procedure on the learning set. This optimum number is used to evaluate the method on the test set.

Tables 5, 6 and 7 summarizes our results on the various data sets in a synthetic way. As we can see, this approach is worth exploring further, as we were able to obtain very good results on some data sets: we are consistently better on the Emotions dataset, and we perform very well on the Yeast dataset for both ranking loss and accuracy. Performances for Scene are worse, but this could be explained by the fact that KNN approaches can perform poorly when a high number of features are used.

[3] The software can be downloaded from https://www.hds.utc.fr/~tdenoeux/.

Table 7 Results for the Emotions dataset

	Worst	Best	BFPC	Rank
F1 (max)	0.431	0.651	0.6673	1
HL (min)	0.361	0.189	0.1988	3
RL (min)	0.331	0.151	0.0907	1
ACC (max)	0.319	0.536	0.5883	1

Roughly speaking, as our approach is based on pairwise information between labels, we can expect to perform better on measures related to pairwise information, such as ranking loss (this is indeed the case for the Yeast data set). Exploring in detail the relation between the current methods and the minimization of some given loss function is out of the scope of the current paper, whose main scope remains making inferences about partial order in a general setting (not necessarily machine learning). This would nevertheless be an interesting study. In any case, empirical results show good performances.

5 Conclusion

In this paper, we have introduced a general uncertainty model, based on belief functions, to deal with partial orders in a pairwise way. Information on each pair of objects (alternatives, labels, ...) regarding ordering or incomparability is modelled by a belief function, and the belief functions of the different pairs are then combined through Dempster's rule. We have then provided means to perform various efficient inference tasks based on the principle of maximal plausibility and on the use of integer linear program.

The application of the approach is then illustrated on two machine learning tasks involving partial orders, namely partial predictions in label ranking and bipartite ranking prediction in multilabel problems. In both cases our approach has proved to be competitive with other algorithms, thus demonstrating its potential interest.

It is worth mentioning that, thanks to the flexibility of our framework, we can in principle predict any partial order. This in contrast with other label ranking methods producing partial orders that can only make predictions that belong to specific families of partial orders, such as semi-orders (Cheng et al. 2012) or interval orders (Destercke 2013).

Given the flexibility of the present approach, it would be interesting to study its connections with other areas such as multi-criteria decision making (for example, we could try to describe the set of orders representable by popular models such a CP-net through constraints). Similarly, it may be worthwhile to explore other potential application domain involving orders, for example object ranking (Kamishima et al. 2011).

Acknowledgments This work was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02).

References

Bache K, Lichman M (2013) UCI machine learning repository. University of California, School of Information and Computer Science, Irvine. http://archive.ics.uci.edu/ml

Blockeel H, Raedt LD, Ramon J (1998) Top-down induction of clustering trees. Proceedings of the 15th international conference on machine learning. Morgan Kaufmann Publishers Inc., San Francisco, pp 55–63

Boutell MR, Luo J, Shen X, Brown CM (2004) Learning multi-label scene classification. Pattern Recogn 37(9):1757–1771

Boutilier C, Brafman RI, Domshlak C, Hoos HH, Poole D (2004) CP-nets: a tool for representing and reasoning with conditional ceteris paribus preference statements. J Artif Intell Res (JAIR) 21:135–191

Cheng W, Rademaker M, De Baets B, Hüllermeier E (2010) Predicting partial orders: ranking with abstention. Machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, pp 215–230

Cheng W, Waegeman W, Welker V, Hüllermeier E (2012) Label ranking with partial abstention based on thresholded probabilistic models. In: Advances in neural information processing systems, pp 2510–2518.

Chow C (1970) On optimum recognition error and reject tradeoff. IEEE Trans Inf Theory 16(1):41–46

Clare A, King RD (2001) Knowledge discovery in multi-label phenotype data. Principles of data mining and knowledge discovery. Springer, New York, pp 42–53

Cobb BR, Shenoy PP (2006) On the plausibility transformation method for translating belief function models to probability models. Int J Approx Reason 41(3):314–330

Dekel O, Singer Y, Manning CD (2003) Log-linear models for label ranking. In: Advances in neural information processing systems, pp 497–504.

Denœux T (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans Syst Man Cybern 25(5):804–813

Denœux T, Masson M-H (2012) Evidential reasoning in large partially ordered sets. Ann Oper Res 195(1):135–161

Destercke S (2013) A pairwise label ranking method with imprecise scores and partial predictions. Machine learning and knowledge discovery in databases. Springer, New York, pp 112–127

El Zoghby N, Cherfaoui V, Denœux T (2013) Optimal object association from pairwise evidential mass functions. 16th international conference on information fusion (FUSION). IEEE, New York, pp 774–780

Elisseeff A, Weston J (2001) A kernel method for multi-labelled classification. In: Advances in neural information processing systems, pp 681–687.

Fürnkranz J, Hüllermeier E (2010) Preference learning. Springer, New York

Fürnkranz J, Hüllermeier E, Mencía EL, Brinker K (2008) Multilabel classification via calibrated label ranking. Mach Learn 73(2):133–153

Grabisch M, Labreuche C (2008) A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. 4OR 6(1):1–44.

Greco S, Kadziński M, SŁowiński R (2011) Selection of a representative value function in robust multiple criteria sorting. Comput Oper Res 38(11):1620–1637

Har-Peled S, Roth D, Zimak D (2003) Constraint classification for multiclass classification and ranking. Adv Neural Inf Process Syst 809–816.

Hüllermeier E, Furnkranz J, Cheng W, Brinker K (2008) Label ranking by learning pairwise preferences. Artif Intell 172(16–17):1897–1916

Kamishima T, Kazawa H, Akaho S (2011) A survey and empirical comparison of object ranking methods. Preference learning. Springer, New York, pp 181–201

Kocev D, Vens C, Struyf J, Džeroski S (2007) Ensembles of multi-objective decision trees. Mach Learn ECML 2007:624–631

Labreuche C (2010) On the robustness for the Choquet integral. Computational intelligence for knowledge-based systems design. Springer, New York, pp 484–493

Li T, Ogihara M (2006) Toward intelligent music information retrieval. IEEE Trans Multimed 8(3):564–574

Loza Mencía E, Park S-H, Fürnkranz J (2010) Efficient voting prediction for pairwise multilabel classification. Neurocomputing 73(7):1164–1176

Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S (2012) An extensive experimental comparison of methods for multi-label learning. Pattern Recogn 45(9):3084–3104

Marden JI (1995) Analyzing and modeling rank data, vol 64. Chapman & Hall, London

Rademaker M, De Baets B (2010) A threshold for majority in the context of aggregating partial order relations. 2010 IEEE international conference on fuzzy systems (FUZZ). IEEE, New York, pp 1–4

Read J, Pfahringer B, Holmes G, Frank E (2011) Classifier chains for multi-label classification. Mach Learn 85(3):333–359

Shafer G (1976) A mathematical theory of evidence. Princeton University Press, New Jersey

Smets P (1993) Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. Int J Approx Reason 9(1):1–35

Smets P, Kennes R (1994) The transferable belief model. Artif Intell 66:191–234

Tritchler D, Lockwood G (1991) Modelling the reliability of paired comparisons. J Math Psychol 35(3):277–293

Troffaes M (2007) Decision making under uncertainty using imprecise probabilities. Int J Approx Reason 45(1):17–29

Tsoumakas G, Katakis I (2007) Multi-label classification: an overview. Int J Data Wareh Min (IJDWM) 3(3):1–13

Tsoumakas G, Katakis I, Vlahavas I (2008) Effective and efficient multilabel classification in domains with large number of labels. In: Proceedings of the ECML/PKDD 2008 workshop on mining multidimensional data (MMD08), pp 30–44.

Tsoumakas G, Katakis I, Vlahavas I (2010) Mining multi-label data. Data mining and knowledge discovery handbook. Springer, New York, pp 667–685

Tsoumakas G, Vlahavas I (2007) Random k-labelsets: an ensemble method for multilabel classification. Machine learning: ECML 2007. Springer, New York, pp 406–417

Ueda N, Saito K (2002) Parametric mixture models for multi-labeled text. In: Advances in neural information processing systems, pp 721–728.

Utkin LV (2009) A new ranking procedure by incomplete pairwise comparisons using preference subsets. Intell Data Anal 13(2):229–241

Vembu S, Gärtner T (2011) Label ranking algorithms: a survey. In: Preference learning, pp 45–64. Springer, New York.

Zaffalon M (2002) The naive credal classifier. J Probab Plan Inference 105:105–122

Zhang M-L, Zhou Z-H (2007) ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048