

Ranking by Pairwise Comparison: A Note on Risk Minimization

Eyke Hüllermeier

Department of Mathematics and Computer Science
Marburg University, Germany
E-mail: eyke@informatik.uni-marburg.de

Johannes Fürnkranz

Department of Computer Science
Darmstadt Technical University, Germany
E-mail: fuernkranz@informatik.tu-darmstadt.de

Abstract—In this paper we consider the problem of learning *ranking functions* in a supervised manner. A ranking function is a mapping from instances to rankings over a finite number of labels and can thus be seen as an extension of a classification function. Our learning method, referred to as *ranking by pairwise comparison* (RPC), is a two-step procedure. First, a valued preference structure is induced from given preference data, using a natural extension of so-called pairwise classification. A ranking is then derived from that preference structure by means of a simple scoring function. It is shown that, under some idealized assumptions, a prediction thus obtained is a risk minimizer if the distance resp. similarity between rankings is measured by the Spearman rank correlation. We conclude the paper by outlining a potential application of the method in (qualitative) fuzzy classification and identifying some extensions necessary in this context.

I. INTRODUCTION

Consider an insurance agent offering his products (insurances) to potential insurees. To be successful (and also to save time), it is important for him to present his offers in the right order. For example, if he knows from experience that middle-aged, working women without children usually prefer offer A to offer B to offer C , he will first suggest A , then maybe B and eventually C . The agent has learned from experience to predict an individual's preferences, expressed in the form of rankings over a set of alternatives, on the basis of features of that individual. This is understood as *preference learning* in this paper.

The problem of preference learning thus defined is particularly challenging for machine learning, as it goes beyond the prediction of single values (such as real numbers in regression analysis and class labels in pattern recognition). In fact, preference learning can be seen as an extension of supervised learning in the classification setting: Instances are not labeled with a single class value but with a set of preferences between pairs of class values.

Needless to say, one might easily think of other types of preference learning. In collaborative filtering, for example, the goal is to predict the products that are liked and might be purchased in the future by a specific user, based on his and other users' previous ratings of products. Here, neither preferences are expressed in terms of rankings, nor individuals characterized by means of feature vectors.

In any case, methods for learning and predicting preferences in an automatic way are among the very recent research topics

in disciplines such as machine learning and recommendation systems. The interest in such methods and the need for computational tools that discover the preferences of individuals is mainly due to an increasing trend toward *personalization* of products and services in e-commerce and various other fields.

The remainder of the paper is organized as follows: In Section II, we briefly recall two standard approaches to preference modeling. In Section III, the problem of preference learning as outlined above is introduced in a formal way, and two rather obvious approaches to learning are indicated. The learning of a ranking function on the basis of valued (fuzzy) preference structures is then discussed in more detail in Section IV. In Section V, it is shown that, under some idealized assumptions, a prediction thus obtained is a risk minimizer if the distance resp. similarity between rankings is measured by the Spearman rank correlation. Finally, we discuss the potential application of our ranking method in the context of (multi-label) classification with (qualitative) fuzzy labels. In this connection, we also suggest some useful extensions of the method and directions for future research.

II. PREFERENCE MODELING

There are two approaches for dealing with preferences that prevail the literature on choice and decision theory. These are based, respectively, on the two perhaps most natural ways for expressing preferences, namely

- by *evaluating* individual alternatives, or
- by *comparing* (pairs of) competing alternatives.

In the latter approach, binary relations are employed in order to express (comparative) preferences in a qualitative way. The basic relation, often denoted $a \succeq b$ or $\mathcal{R}(a, b)$, is usually interpreted as “alternative a is at least as good as alternative b ”. The reflexive relation $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$, where \mathcal{A} is the set of alternatives, induces a strict preference relation \mathcal{P} , an indifference relation \mathcal{I} and an incomparability relation \mathcal{J} in a straightforward way. A triple $(\mathcal{P}, \mathcal{I}, \mathcal{J})$ with an asymmetric relation \mathcal{P} , a reflexive and symmetric relation \mathcal{I} , and a symmetric relation \mathcal{J} is often referred to as a *preference structure*.¹

¹Formally, $(\mathcal{P}, \mathcal{I}, \mathcal{J})$ must also satisfy the following: $\mathcal{P} \cap \mathcal{I} = \emptyset$, $\mathcal{P} \cap \mathcal{J} = \emptyset$, $\mathcal{I} \cap \mathcal{J} = \emptyset$, $\mathcal{P} \cup \mathcal{I} \cup \mathcal{J} = \mathcal{A} \times \mathcal{A}$.

The second approach is more inclined to numerical representations of preferences via *utility* or *value functions*. A utility function f assigns an abstract utility degree to each alternative under consideration and thus induces a preference relation \succeq by virtue of $\mathcal{R}(a, b) \Leftrightarrow f(a) \geq f(b)$. The relationship between the two approaches has been one of the foci of research in (multi-attribute) utility theory. A question of interest concerns, e.g., the characterization of preference structures that can be represented by utility functions having a certain (simple) mathematical structure.²

Obviously, the numerical approach provides stronger information than the relational one but is also more restrictive and more demanding from a modeling point of view. In fact, it is usually much easier for people to provide *relative/comparative* preference information of the form “I prefer alternative a to alternative b ” than *absolute* information of the type “I like alternative a to the degree 0.9”, requiring the specification of a utility degree (score) for each alternative.

Moreover, *revealed/observed* preferences are usually of the relational type. Consider, for instance, a person in a situation where a choice between two alternatives a and b must be made. If that person chooses a and not b , it is at least likely that he prefers a to b . Formally, this gives rise to the information $\mathcal{R}(a, b)$, whereas nothing is known about the *absolute* utility degrees for alternatives a and b .

Recent advances in both approaches to preference modeling have become possible through concepts and tools from fuzzy set theory. *Valued* or *fuzzy* preference structures extend the classical approach by replacing the binary relation \mathcal{R} with a fuzzy relation: $\mathcal{R}(a, b)$ is no longer forced to be either 0 (which means that $a \not\succeq b$) or 1 (which means that $a \succeq b$) but can take any value in the unit interval $[0, 1]$ (or, more generally, some linearly ordered scale). The value $\mathcal{R}(a, b)$ can be interpreted mainly in two ways, namely (i) as a degree of *intensity* of the preference of a over b [3], or (ii) as a degree of uncertainty or confidence of that preference [10]. An associated fuzzy preference structure is a triple $(\mathcal{P}, \mathcal{I}, \mathcal{J})$ of fuzzy relations having certain mathematical properties which are generalizations of the properties required in the non-fuzzy case.³ Fuzzy preference structures, especially their axiomatic construction, have been studied extensively in literature [1].

III. PREFERENCE LEARNING

A. The Learning Problem

In order to present the problem of preference learning, as considered in this paper, in a formal way, let \mathcal{X} denote a set of instances characterized in terms of an attribute-value representation: $\mathcal{X} = X_1 \times X_2 \times \dots \times X_l$, where X_i is the domain of the i -th attribute. Thus, an instance is represented as a vector $x = (x_1 \dots x_l) \in \mathcal{X}$. Moreover, let $\mathcal{L} = \{\lambda_1 \dots \lambda_m\}$ be a set of alternatives/labels. We assume that each instance

has preferences concerning the alternatives $\lambda \in \mathcal{L}$. Formally, this can be expressed in terms of a *preference function*

$$\mathcal{X} \rightarrow \mathfrak{P}, \quad (1)$$

which maps instances to preference models; \mathfrak{P} denotes the class of potential models.

Note that an instance $x \in \mathcal{X}$ does not necessarily represent a single individual. In general, it may represent a set of several individuals. For example, if people are characterized in terms of a few attributes such as, e.g., sex and age, it is of course possible that two people are mapped to the same representation. Thus, even under the assumption that the preferences of a single individual are fixed, one should actually associate a probability distribution over \mathfrak{P} with each instance x rather than a single model. In this sense, (1) should be considered as a mapping that assigns a *representative* model to each instance, i.e., a model that is representative of an underlying distribution. We shall come back to this point in Section V below.

The problem of *preference learning* now consists of learning (approximating) the mapping (1) on the basis of information about the preferences of instances (examples) x^k , $k = 1 \dots n$. More specifically, we assume that a single piece of information corresponds to *comparative* preference information of the form

$$\lambda_i \succ_{x^k} \lambda_j, \quad (2)$$

i.e. “individual x^k prefers λ_i to λ_j ”. As already mentioned above, this type of information is often easier to obtain than absolute ratings of single alternatives in terms of utility degrees.

Regarding the class of preference models \mathfrak{P} , we are especially interested in *rankings* of alternatives. For convenience, we shall subsequently assume that these rankings are not only weak but even total orders (i.e. no ties are allowed in the ordering). The ranking \succ_x of an individual x can thus be expressed in terms of a permutation τ_x of $\{1 \dots m\}$ such that

$$\lambda_{\tau_x(1)} \succ_x \lambda_{\tau_x(2)} \succ_x \dots \succ_x \lambda_{\tau_x(m)}. \quad (3)$$

Here, $\lambda_i \succ_x \lambda_j$ means that x (strictly) prefers λ_i to λ_j . The relation between \succ_x and τ_x is obviously one-to-one. Subsequently, we shall employ both notations. We denote the class of all permutations of $\{1 \dots m\}$ by \mathcal{S}_m .

Note that knowledge about a complete ranking (3) can be expanded into $m(m-1)/2$ comparative preferences $\lambda_{\tau_x(i)} \succ \lambda_{\tau_x(j)}$ of the form (2).

B. Approaches to Preference Learning

In the previous section, two natural approaches to preference modeling have been outlined. Correspondingly, there are two natural approaches to preference learning or, more precisely, to the learning of rankings.

A first rather obvious idea is to estimate a kind of utility degree for each alternative λ_i and to rank order the alternatives according to these utility degrees. This approach can be seen as a straightforward extension of the *winner-takes-all* strategy for multi-class classification. Here, a real-valued function f_i is

²Of special interest are utility functions that can be decomposed into additive *subutility* functions.

³The question of how to generalize these properties is non-trivial. Especially, there are different options which are formally not equivalent.

associated with each class label λ_i and a query instance x is assigned to the class for which $f_i(x)$ is maximal. In the context of ranking, the permutation $\hat{\tau}_x$ estimated for an instance x could thus be expressed as follows:

$$\hat{\tau}_x = \operatorname{argsort}\{f_i(x) \mid i = 1 \dots m\}, \quad (4)$$

where $\operatorname{argsort}$ returns a permutation τ of $\{1 \dots m\}$ such that $\tau(i) \leq \tau(j)$ if $f_i(x) \geq f_j(x)$.⁴

With regard to this learning method it should be noted that, according to our assumption, the utility degrees $f_i(x^k)$ are actually not available for the examples x^k . Rather, only comparative preference information of the form (2) is given. Still, a preference $\lambda_i \succ_{x^k} \lambda_j$ can obviously be turned into a *constraint* of the form $f_i(x^k) > f_j(x^k)$, and from such constraints the functions $f_1 \dots f_m$ might eventually be estimated. An approach of this kind is pursued in [8].

A second idea, which is in line with the relational approach to preference modeling, is to predict the *pairwise* preferences for a query instance x , i.e., the relation \mathcal{R}_x .⁵ Since the empirical data just consists of such preferences, this idea might be considered even more natural. Nevertheless, a collection of estimated pairwise preferences does not necessarily determine a ranking in an unequivocal way. Therefore, an additional *ranking procedure* is needed, which turns a collection of pairwise preferences into a ranking. This approach to preference learning will be discussed in more detail in the following section.

IV. PAIRWISE PREFERENCE LEARNING

The idea of pairwise learning is well-known in the context of classification [6], where it allows one to transform an m -class classification problem, i.e., a problem involving $m > 2$ classes $\mathcal{L} = \{\lambda_1 \dots \lambda_m\}$, into a number of *binary* problems. To this end, a separate model (base learner) \mathcal{M}_{ij} is trained for each *pair* of labels $(\lambda_i, \lambda_j) \in \mathcal{L}$, $1 \leq i < j \leq m$. Thus, a total number of $m(m-1)/2$ models is needed. \mathcal{M}_{ij} is intended to separate the classes C_i (instances with label λ_i) and C_j .

At classification time, a query is submitted to all learners, and each prediction is interpreted as a vote for a label: If classifier \mathcal{M}_{ij} predicts λ_i , this is counted as a vote for λ_i , just as the prediction λ_j would be considered as a vote for λ_j . The label with the highest number of votes is then proposed as a prediction.

The above procedure can be extended to the case of preference learning or, more precisely, the learning of rankings in a natural way [7]. A preference information of the form $\lambda_i \succ_x \lambda_j$ is turned into a training example (x, y) for the learner \mathcal{M}_{ab} , where $a = \min(i, j)$ and $b = \max(i, j)$. Moreover, $y = 1$ if $i < j$ and $= 0$ otherwise. Thus, \mathcal{M}_{ab} is intended to learn the mapping

$$x \mapsto \begin{cases} 1 & \text{if } \lambda_a \succ_x \lambda_b \\ 0 & \text{if } \lambda_b \succ_x \lambda_a \end{cases}. \quad (5)$$

⁴Any rule can be used for tie breaking.

⁵Note that $\mathcal{R} = \mathcal{P}$ in the case where $\mathcal{I} = \emptyset$.

In other words, given an instance x as an input, \mathcal{M}_{ab} is assumed to output 1 if $\lambda_a \succ_x \lambda_b$ and 0 if $\lambda_b \succ_x \lambda_a$.

The mapping (5) can be realized by any binary classifier. Alternatively, one might of course also employ a classifier that maps into $[0, 1]$ instead of $\{0, 1\}$. The output of such a “soft” binary classifier can usually be interpreted as a probability or, more generally, a kind of confidence in the classification. Thus, the closer the output of \mathcal{M}_{ab} to 1, the stronger the preference $\lambda_a \succ_x \lambda_b$ is supported.

A soft classifier naturally leads to a valued (fuzzy) preference relation \mathcal{R}_x associated with an instance x :

$$\mathcal{R}_x(\lambda_i, \lambda_j) = \begin{cases} \mathcal{M}_{ij}(x) & \text{if } i < j \\ 1 - \mathcal{M}_{ij}(x) & \text{if } i > j \end{cases} \quad (6)$$

for all $\lambda_i \neq \lambda_j \in \mathcal{L}$. Thus, we have obtained a preference learner, composed of an ensemble of (soft) binary classifiers, which can be constructed on the basis of training data in the form of individuals with associated (partial) preferences. This preference learner assigns a valued preference relation to any (query) instance $x \in \mathcal{X}$.

Let us now return to the problem of predicting the ranking of an instance x , characterized in terms of a permutation τ_x . As we mentioned above, a relation \mathcal{R}_x does not always suggest a unique ranking in an unequivocal way. In fact, the problem of inducing a ranking from a (valued) preference relation has received a lot of attention in several research fields, e.g., in fuzzy preference modeling and (multi-attribute) decision making [5].

A straightforward extension of the aforementioned voting strategy for classification is to evaluate the alternatives λ_i by means of the sum of (weighted) votes

$$S(\lambda_i) = \sum_{\lambda_j \neq \lambda_i} \mathcal{R}_x(\lambda_i, \lambda_j) \quad (7)$$

and to rank order them according to these evaluations, i.e.

$$(\lambda_i \succeq \lambda_j) \Leftrightarrow (S(\lambda_i) \geq S(\lambda_j)). \quad (8)$$

This is a particular type of “ranking by scoring” strategy, where the alternatives are assigned a score first and then ordered according to these scores. Here, the scoring function given by (7).

The combination of learning a preference relation (6) and applying the voting procedure (7) will be referred to as *ranking by pairwise comparison* (RPC). Note that RPC is in a sense very similar to the alternative approach (4) to preference learning as outlined in the previous section. In fact, the scores $S(\lambda_i)$ in (7) can directly be compared with the values $f_i(x)$ in (4): In both approaches, a score for each alternative λ_i is eventually derived, even if this is done in an *indirect* way in RPC. On the one hand, one might of course argue that an indirect approach is unnecessarily complex and maybe less effective. On the other hand, it should be noted that the individual learning problems faced by the base learners \mathcal{M}_{ij} are usually less difficult than the problem of learning the functions $f_i(\cdot)$ directly. Roughly speaking, this is due to the fact that a learner \mathcal{M}_{ij} has to separate only two labels. As

opposed to this, all the $f_i(\cdot)$ have to be consistent among each other in the direct approach. In [8], for example, the problem of learning a ranking function is also reduced to a standard classification problem. However, since in this approach the functions $f_i(\cdot)$ are learned simultaneously, the classification problem becomes extremely high-dimensional. Comparing the pros and cons of the two approaches, namely learning and combining several low-dimensional classifiers versus learning one high-dimensional classifier, is a topic of ongoing research.

V. RPC AND RISK MINIMIZATION

The quality of a model \mathcal{M} (induced by a learning algorithm) is commonly expressed in terms of its *expected loss* or *risk*

$$\mathbb{E} (D(y, \mathcal{M}(x))), \quad (9)$$

where $D(\cdot)$ is a loss or distance function, $\mathcal{M}(x)$ denotes the prediction made by the learning algorithm for the instance x , and y is the true outcome. The expectation \mathbb{E} is taken over $\mathcal{X} \times \mathcal{Y}$, where \mathcal{Y} is the output space (e.g. the set \mathcal{L} of classes in classification).⁶

A. 0/1-Loss

In the case of a 0/1-loss function, i.e., $D(y, \hat{y}) = 0$ for $y = \hat{y}$ and $= 1$ otherwise, the optimal (Bayes) prediction for a specific instance x is obviously given by the most probable outcome y . In the classification setting, for example, where $\mathcal{Y} = \mathcal{L}$, this estimate is the class with maximum posterior probability $\Pr(\lambda_i | x)$. A straightforward generalization of this principle to the ranking setting, where \mathcal{Y} is the class of rankings over \mathcal{L} , leads to the prediction

$$\hat{\tau}_x = \arg \max_{\tau \in \mathcal{S}_k} \Pr(\tau | x), \quad (10)$$

where $\Pr(\tau | x)$ is the conditional probability of a ranking (permutation) given an instance x .

Unfortunately, (10) cannot be derived by the RPC approach, a problem that is mainly caused by a loss of information about the complete rankings of the training examples: Recall that even if complete rankings are available for training, these rankings are first split into pairwise preferences, which are then used for training the base learners. Now, even though the probabilities $\Pr(\lambda_i \succ_x \lambda_j)$ of pairwise preferences can obviously be derived from the distribution $\Pr(\cdot | x)$, the latter cannot be recovered from the former. To illustrate, consider the following distributions $\Pr(\cdot | x) \neq \Pr'(\cdot | x)$:

$\tau(1)$	$\tau(2)$	$\tau(3)$	$\Pr(\tau x)$	$\Pr'(\tau x)$
1	2	3	0.5	0.5
1	3	2	0.1	0
2	1	3	0	0.1
3	1	2	0.4	0.3
2	3	1	0	0.1
3	2	1	0	0

From both distributions, one derives the same probabilities $\Pr(\lambda_i \succ_x \lambda_j) = \sum_{\tau: \tau(i) < \tau(j)} \Pr(\tau|x)$:

⁶The existence of a probability measure over $\mathcal{X} \times \mathcal{Y}$ must of course be assumed.

	λ_1	λ_2	λ_3
λ_1	–	0.6	0.6
λ_2	0.4	–	0.9
λ_3	0.4	0.1	–

Moreover, in the (ideal) case where the base learners' estimates correspond exactly to the probabilities of pairwise preference, i.e.,

$$\mathcal{R}_x(\lambda_i, \lambda_j) = \mathcal{M}_{i,j}(x) = \Pr(\lambda_i \succ_x \lambda_j), \quad (11)$$

the prediction obtained by RPC is $\lambda_2 \succ_x \lambda_1 \succ_x \lambda_3$. This prediction has a very low probability (even 0 in the case of \Pr) and obviously differs from the ranking that is optimal in the sense of (10), namely $\lambda_1 \succ_x \lambda_2 \succ_x \lambda_3$.

B. Pairwise Ranking and Rank Correlation

Needless to say, the simple 0/1-distance function is a rather questionable measure for rankings, as it does not take into account that two rankings can be more or less similar. And indeed, a number of more sophisticated distance measures for rankings have been proposed in literature. As we will show now, RPC yields a risk minimizing prediction

$$\hat{\tau}_x = \arg \min_{\tau \in \mathcal{S}_k} \sum_{\tau' \in \mathcal{S}_m} D(\tau, \tau') \cdot \Pr(\tau' | x) \quad (12)$$

for an important and frequently applied distance measure $D(\cdot)$, namely for the sum of squared rank distances

$$D(\tau', \tau) =_{\text{def}} \sum_{i=1}^m (\tau'(i) - \tau(i))^2 \quad (13)$$

and, hence, for the *Spearman rank correlation*⁷

$$1 - \frac{6D(\tau, \tau')}{m(m^2 - 1)} \in [-1, 1]. \quad (14)$$

Lemma 1: Let $s_i, i = 1 \dots m$, be real numbers such that $0 \leq s_1 \leq s_2 \dots \leq s_m$. Then, for all permutations $\tau \in \mathcal{S}_m$,

$$\sum_{i=1}^m (i - s_i)^2 \leq \sum_{i=1}^m (i - s_{\tau(i)})^2 \quad (15)$$

Proof: We have

$$\begin{aligned} \sum_{i=1}^m (i - s_{\tau(i)})^2 &= \sum_{i=1}^m (i - s_i + s_i - s_{\tau(i)})^2 \\ &= \sum_{i=1}^m (i - s_i)^2 + 2 \sum_{i=1}^m (i - s_i)(s_i - s_{\tau(i)}) \\ &\quad + \sum_{i=1}^m (s_i - s_{\tau(i)})^2. \end{aligned}$$

Expanding the last equation and exploiting that $\sum_{i=1}^m s_i^2 = \sum_{i=1}^m s_{\tau(i)}^2$ yields

$$\sum_{i=1}^m (i - s_{\tau(i)})^2 = \sum_{i=1}^m (i - s_i)^2 + 2 \sum_{i=1}^m i s_i - 2 \sum_{i=1}^m i s_{\tau(i)}.$$

⁷This is of course a similarity rather than a distance measure.

On the right-hand side of the last equation, only the last term $\sum_{i=1}^m i s_{\tau(i)}$ depends on τ . Since $s_i \leq s_j$ for $i < j$, this term becomes maximal for $\tau(i) = i$. Therefore, the right-hand side is larger than or equal to $\sum_{i=1}^m (i - s_i)^2$, which proves the lemma. \square

Lemma 2: Let $\Pr(\cdot | x)$ be a probability distribution over \mathcal{S}_m and let $p(\tau) =_{\text{def}} \Pr(\tau | x)$. Moreover, let

$$s_i =_{\text{def}} m - \sum_{j \neq i} \Pr(\lambda_i \succ_x \lambda_j) \quad (16)$$

with

$$\Pr(\lambda_i \succ_x \lambda_j) = \sum_{\tau: \tau(j) < \tau(i)} \Pr(\tau | x). \quad (17)$$

Then, $s_i = \sum_{j \neq i} p(\tau) \tau(i)$.

Proof: We have

$$\begin{aligned} s_i &= m - \sum_{j \neq i} \Pr(\lambda_i \succ_x \lambda_j) \\ &= 1 + \sum_{j \neq i} (1 - \Pr(\lambda_i \succ_x \lambda_j)) \\ &= 1 + \sum_{j \neq i} \Pr(\lambda_j \succ_x \lambda_i) \\ &= 1 + \sum_{j \neq i} \sum_{\tau: \tau(j) < \tau(i)} p(\tau) \\ &= 1 + \sum_{\tau} p(\tau) \sum_{j \neq i} \begin{cases} 1 & \text{if } \tau(i) > \tau(j) \\ 0 & \text{if } \tau(i) < \tau(j) \end{cases} \\ &= 1 + \sum_{\tau} p(\tau) (\tau(i) - 1) \\ &= \sum_{\tau} p(\tau) \tau(i) \end{aligned}$$

Remark: Note that $s_i \leq s_j$ is equivalent to $S(\lambda_i) \geq S(\lambda_j)$ under the assumption (11). Thus, ranking the alternatives according to $S(\lambda_i)$ (in decreasing order) is equivalent to ranking them according to s_i (in increasing order).

Theorem 3: The expected distance

$$\begin{aligned} E(\tau') &= \sum_{\tau} p(\tau) \cdot D(\tau', \tau) \\ &= \sum_{\tau} p(\tau) \sum_{i=1}^m (\tau'(i) - \tau(i))^2 \end{aligned} \quad (18)$$

becomes minimal by choosing τ' such that $\tau'(i) \leq \tau'(j)$ whenever $s_i \leq s_j$, with s_i given by (16).

Proof: We have

$$\begin{aligned} E(\tau'_x) &= \sum_{\tau} p(\tau) \sum_{i=1}^m (\tau'_x(i) - \tau(i))^2 \\ &= \sum_{i=1}^m \sum_{\tau} p(\tau) (\tau'_x(i) - \tau(i))^2 \\ &= \sum_{i=1}^m \sum_{\tau} p(\tau) (\tau'_x(i) - s_i + s_i - \tau(i))^2 \\ &= \sum_{i=1}^m \sum_{\tau} p(\tau) [(\tau(i) - s_i)^2 - 2(\tau(i) - s_i)(s_i - \tau'(i)) \\ &\quad + (s_i - \tau'(i))^2] \\ &= \sum_{i=1}^m \left[\sum_{\tau} p(\tau) (\tau(i) - s_i)^2 - 2(s_i - \tau'(i)) \cdot \right. \\ &\quad \left. \cdot \sum_{\tau} p(\tau) (\tau(i) - s_i) + \sum_{\tau} p(\tau) (s_i - \tau'(i))^2 \right] \end{aligned}$$

In the last equation, the mid-term on the right-hand side becomes 0 according to Lemma 2. Moreover, the last term obviously simplifies to $(s_i - \tau'(i))$, and the first term is a constant $c = \sum_{\tau} p(\tau) (\tau(i) - s_i)^2$ that does not depend on τ' . Thus, we obtain $E(\tau'_x) = c + \sum_{i=1}^m (s_i - \tau'(i))^2$ and the theorem follows from Lemma 1. \square

C. Connections with Voting Theory

It is worth mentioning that RPC is closely related to the so-called *Borda-count*, a voting rule that is well-known in social choice theory: Suppose that the preferences of n voters are expressed in terms of rankings $\tau_1, \tau_2 \dots \tau_n$ of m alternatives. From a ranking τ_i , the following scores are derived for the alternatives: The best alternative receives $m - 1$ points, the second best $m - 2$ points, and so on. The overall score of an alternative is the sum of points that it has received from all voters, and a representative ranking $\hat{\tau}$ (aggregation of the single voters' rankings) is obtained by rank ordering the alternatives according to these scores.

Now, it is readily verified that the result obtained by this procedure corresponds exactly to the result of RPC if the probability distribution over the class \mathcal{S}_m of rankings is defined by the corresponding relative frequencies. In other words, the ranking $\hat{\tau}$ obtained by RPC minimizes the sum of all distances:

$$\hat{\tau} = \arg \min_{\tau \in \mathcal{S}_k} \sum_{i=1}^n D(\tau, \tau_i). \quad (19)$$

In connection with social choice theory it is also interesting to note that RPC does not satisfy the so-called *Condorcet criterion*: As the pairwise preferences in our above example show, it is thoroughly possible that an alternative (in this case λ_1) is preferred in all *pairwise* comparisons ($\mathcal{R}(\lambda_1, \lambda_2) > .5$ and $\mathcal{R}(\lambda_1, \lambda_3) > .5$) without being the overall winner of the election. Of course, this apparently paradoxical property is not only relevant for ranking but also for classification. In this context, it has already been recognized in [9].

A distance measure for rankings, which plays an important role in voting theory, is the so-called *Kendall's tau*. This measure is defined by the minimal number of pairwise inversions of (adjacent) labels needed to transform the first ranking into the second one. When using Kendall's tau as a distance measure $D(\cdot)$ in (19), $\hat{\tau}$ is also called the *Kemeny-optimal* ranking. Kendall's tau is intuitively quite appealing and Kemeny-optimal rankings have several nice properties. However, one drawback of using Kendall's tau instead of rank correlation as a distance measure in (19) is a loss of computational efficiency. In fact, the computation of Kemeny-optimal rankings is known to be NP-hard [2].

VI. CLASSIFICATION WITH FUZZY LABELS

So far, the relation \succ has always been interpreted as "is preferred to", i.e., as a preference relation in the narrow sense. However, one can easily imagine other interpretations and, hence, other applications of ranking (by pairwise comparison).

For example, consider again the standard classification problem. Quite often, an instance x cannot be assigned to one class in an unequivocal way. In that case, the potential class labels might be *rank ordered* according to their suitability: A label λ_i is "preferred" to a label λ_j if it appears more suitable for x . For example, *middle-aged* \succ_x *young* \succ_x *old* would then mean that a person x is best characterized as *middle-aged*, then as *young*, and least as *old*.

Such a ranking of labels can be considered as a qualitative version of a *fuzzy label*. The latter is commonly understood as a fuzzy subset of the set of labels \mathcal{L} , i.e., as a mapping $\mu_x : \mathcal{L} \rightarrow [0, 1]$ such that $\mu_x(\lambda_i)$ is the degree to which the label λ_i applies to the instance x resp. the degree of possibility that x belongs to class λ_i [4]. Assigning a qualitative fuzzy label to an instance by expressing pairwise "preferences" will often be less difficult for an expert than assigning a possibility distribution in the form of real numbers $\mu_x(\lambda_1), \mu_x(\lambda_2) \dots \mu_x(\lambda_m)$. Thus, applying our RPC approach might be of great practical relevance for classification problems.

However, this type of application suggests some extensions of RPC as presented in this paper. In fact, it should be noticed that a qualitative fuzzy label is not necessarily a complete ranking. Usually, there are several labels that do definitely not apply to the instance under consideration. For example, suppose that λ_1 appears completely appropriate, λ_2 not so much but still more than the remaining labels $\lambda_3 \dots \lambda_m$. This can be expressed by the following set of pairwise preferences: $\lambda_1 \succ_x \lambda_i$ for $i = 2 \dots m$ and $\lambda_2 \succ_x \lambda_j$ for $j = 3 \dots m$. Thus, modeling qualitative fuzzy labels in terms of a ranking calls for two extensions. Firstly, the *ex aequo* case $\lambda_i \sim \lambda_j$ should not be excluded, i.e., it should be possible to express indifference between two labels. Secondly, a ranking is naturally split into two parts, namely the "positive" labels that apply to the instance x at least to some extent, and the labels that have a more "negative" association. From a classification point of view, it would of course be useful to endow a ranking with information about this split.

Regarding the first point, the problem might be solved by training separate models $\mathcal{M}_{i,j}$ and $\mathcal{M}_{j,i}$. In the binary (non-fuzzy) case, the relation between the labels λ_i and λ_j can then be recovered from $\mathcal{M}_{i,j}$'s vote u and $\mathcal{M}_{j,i}$'s vote v as follows: $\lambda_i \succ \lambda_j$ if $(u, v) = (1, 0)$, $\lambda_i \prec \lambda_j$ if $(u, v) = (0, 1)$, $\lambda_i \sim \lambda_j$ if $(u, v) = (1, 1)$. (Note that in practice the case $u = v = 0$ can occur. This indicates a conflict and means that (at least) one learner has made an incorrect vote.) Regarding the second point, it is clear that the required extension cannot be deduced from *pairwise* comparisons but has to be learned separately. Interestingly enough, this extension is closely related to the idea of a *denoted* utility scale [11], i.e., a "bipolar" scale that consists of a positive part and a negative part (with a denoted element in-between). Thus, in our context we may speak of a "denoted ranking" that might look as in the following example: $\lambda_2 \succ \lambda_4 \sim \lambda_3 \mid \lambda_5 \sim \lambda_1 \succ \lambda_6$. Here, the symbol \mid indicates the separation between positive and negative labels. Note that it automatically implies a strict preference, i.e. $\lambda_3 \succ \lambda_5$ in the example above.

One possibility to approach the problem of learning a denoted ranking is to complement the RPC procedure with a standard multi-label classification: The positive labels ($\lambda_2, \lambda_4, \lambda_3$ in the above example) are the labels that apply to the instance x , whereas the negative ones do not apply.

VII. CONCLUSION

We have presented a method for learning ranking functions that is in line with the relational approach to preference modeling. As an important property of our method we have shown that it yields risk minimizing predictions if the quality of a prediction is expressed in terms of (Spearman) rank correlation. Finally, we have indicated the applicability of the method in the context of fuzzy classification. Developing this idea in more detail is an important aspect of future work.

REFERENCES

- [1] B De Baets and J Fodor. Twenty years of fuzzy preference structures (1978-1997). *Riv. Mat. Sci. Econom. Social.*, 20:45–66, 1997.
- [2] JJ Bartholdi, CA Tovey, and MA Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6(2), 1989.
- [3] J Bezdek, B Spillman, and R Spillman. A fuzzy relation for group decision theory. *Fuzzy Sets and Systems*, 1:255–268, 1978.
- [4] T Denoeux and LM Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122:47–62, 2001.
- [5] J Fodor and M. Roubens. *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer, 1994.
- [6] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.
- [7] J. Fürnkranz and E. Hüllermeier. Pairwise preference learning and ranking. In ECML–2003, *Proceedings 13th European Conference on Machine Learning*, Cavtat-Dubrovnik, Croatia, September 2003.
- [8] S. Har-Peled, D. Roth, and D. Zimak. Constraint classification: a new approach to multiclass classification. In *Proceedings 13th Int. Conf. on Algorithmic Learning Theory*, pages 365–379, Lübeck, Germany, 2002.
- [9] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.
- [10] S. Orlovski. Decision-making with a fuzzy preference relation. *Fuzzy Sets and Systems*, 1:155–167, 1978.
- [11] R.R. Yager. Using a notion of acceptable in uncertain ordinal decision making. *Int. J. Uncert., Fuzziness, and Knowl.-Based Syst.*, 10(3), 2002.