Preference Learning Using the Choquet Integral: The Case of Multipartite Ranking

Ali Fallah Tehrani, Weiwei Cheng, and Eyke Hüllermeier

***Abstract*—We propose a novel method for preference learning or, more specifically, learning to rank, where the task is to learn a ranking model that takes a subset of alternatives as input and produces a ranking of these alternatives as output. Just like in the case of conventional classifier learning, training information is provided in the form of a set of labeled instances, with labels or, say, preference degrees taken from an ordered categorical scale. This setting is known as multipartite ranking in the literature. Our approach is based on the idea of using the (discrete) Choquet integral as an underlying model for representing ranking functions. Being an established aggregation function in fields such as multiple criteria decision making and information fusion, the Choquet integral offers a number of interesting properties that make it attractive from a machine learning perspective, too. The learning problem itself comes down to properly specifying the fuzzy measure on which the Choquet integral is defined. This problem is formalized as a margin maximization problem and solved by means of a cutting plane algorithm. The performance of our method is tested on a number of benchmark datasets.**

***Index Terms*—Attribute interactions, Choquet integral, classification, monotonicity, preference learning.**

I. Introduction

PREFERENCE learning is an emerging subfield of machine learning that has received increasing attention in recent years [1]. Roughly speaking, the goal in preference learning is to induce preference models from observed data that reveals information about the preferences of an individual or a group of individuals in a direct or indirect way; these models are then used to predict the preferences in a new situation. In this regard, predictions in the form of *rankings*, i.e., total orders of a set of alternatives, constitute an important special case [2]–[6]. A ranking can be seen as a specific type of *structured output* [7], and compared to conventional classification and regression functions, models producing such outputs require a more complex internal representation.

In this paper, we propose a novel method for such kind of ranking problems, for the first time using the (discrete) Choquet integral [8] as an underlying model for representing rankings in a setting of supervised learning. The Choquet integral is an established aggregation function that has been used in various fields of application, including multiple criteria decision making and information fusion. It can be seen as a generalization of the weighted arithmetic mean that is not only able to capture the importance of individual features but also information about the interaction (e.g., redundancy or complementarity) between different features. Moreover, it obeys monotonicity properties in a rather natural way. Due to these properties, the Choquet integral appears to be very appealing for preference learning, especially for aggregating the evaluation of individual features in the form of interacting criteria. The learning problem itself comes down to specifying the fuzzy measure underlying the definition of the Choquet integral in the most suitable way. In this regard, we explore connections to kernel-based machine learning methods [9].

While a number of different types of ranking problems have been introduced in the literature in recent years, we specifically focus on a setting referred to as *multipartite ranking* [2], [6]. Roughly speaking, the task in multipartite ranking is to learn a ranking model that takes any set of alternatives as input, with each alternative typically represented in terms of a feature vector, and produces a ranking of these alternatives as output. Just like in the case of conventional classifier learning, training information is provided in the form of a set of labeled instances, with labels or, say, preference degrees taken from an ordered categorical scale (such as bad, good, and very good).

The remainder of this paper is organized as follows. In the next section, we give a brief overview of related work. In Section III, we recall the basic definition of the (discrete) Choquet integral and related notions. The ranking problem we are dealing with and our method for tackling this problem are introduced in Sections IV and V, respectively. Experimental results are presented in Section VI, prior to concluding this paper in Section VII.

II. Related Work

In this section, we briefly review related work on preference learning and the use of the Choquet integral in machine learning.

A. Preference Learning

Methods for the automatic learning, discovery, and adaptation of preferences have received increasing interest in machine learning, data mining, and related research fields in recent years. Approaches relevant to this area range from preference elicitation where the utility function of a single agent is estimated by asking questions effectively [10]–[12] to *collaborative filtering* where a customer's preferences are estimated from the preferences of other customers [13], [14]. Preference learning can be formalized within various settings, depending, e.g., on the

Manuscript received August 10, 2011; revised January 8, 2012; accepted February 12, 2012. Date of publication April 26, 2012; date of current version November 27, 2012.

The authors are with the Department of Mathematics and Computer Science, University of Marburg, Marburg 35032, Germany (e-mail: fallah@mathematik.uni-marburg.de; cheng@mathematik.uni-marburg.de; eyke@mathematik.uni-marburg.de).

Digital Object Identifier 10.1109/TFUZZ.2012.2196050

underlying preference model and the type of information provided as an input to the learning system.

There are two main approaches to modeling preferences that prevail in the literature on choice and decision theory: value functions and preference relations. From a machine learning point of view, these two approaches give rise to two kinds of learning problems: learning value functions and learning (binary) preference relations. The latter deviates more strongly than the former from conventional problems such as classification and regression, as it involves the prediction of complex structures, such as rankings or partial order relations, rather than single values. Moreover, training input in preference learning will not, as it is usually the case in supervised learning, be offered in the form of complete examples but may comprise more general types of information, such as relative preferences or different kinds of indirect feedback and implicit preference information [15], [16].

In general, a preference learning system is provided with a set of items (e.g., products) for which preferences are known, and the task is to learn a function that predicts preferences for a new set of items (e.g., new products not seen so far), or for the same set of items in a different context (e.g., the same products but for a different user). Frequently, the predicted preference relation is required to form a total order, in which case we also speak of a *ranking problem*. In fact, among the problems in the realm of preference learning, the task of "learning to rank" has probably received the most attention in the literature so far, and a number of different ranking problems have already been introduced. Based on the type of training data and the required predictions, Fürnkranz and Hüllermeier [1] distinguish between the problems of object ranking [2], [17], label ranking [18]–[21], and instance ranking [6], [22].

All of these basic learning tasks can be tackled by similar techniques. As with the distinction between using value functions and binary relations for modeling preferences, two general approaches to preference learning have been proposed in the literature, the first one being based on the idea of learning to evaluate individual alternatives by means of a value function, while the second one seeks to compare (pairs of) competing alternatives, that is, to learn one or more binary preference predicates. Making sufficiently restrictive model assumptions about the structure of a preference relation, one can also try to use the data for identifying this structure. Finally, local estimation techniques such as nearest neighbor prediction can be used, which mostly lead to aggregating preferences in one way or another.

A value function assigns an abstract degree of utility to each alternative under consideration. Depending on the underlying utility scale, which is typically either numerical or ordinal, the problem of learning a (latent) value function from the given training data becomes one of regression learning or ordinal classification. Both problems are well known in machine learning. However, value functions often implicate special requirements and constraints that have to be taken into consideration such as, for example, monotonicity in certain attributes. Besides, as mentioned earlier, training data are not necessarily given in the form of input/output pairs, i.e., alternatives (instances) together with their utility degrees, but may also consist of qualitative feedback in the form of pairwise comparisons, stating that one alternative is preferred to another one and therefore has a higher utility degree. In general, this means that value functions need to be learned from indirect instead of direct training information [15], [16].

The learning of binary preference relations that compare alternatives in a pairwise manner is normally simpler, mainly because comparative training information (suggesting that one alternative is better than another one) can be used directly instead of translating it into constraints on a (latent) value function [4], [23]. On the other hand, the prediction step may become more difficult, since a binary preference relation learned from data is not necessarily consistent in the sense of being transitive and, therefore, does normally not define a ranking in a unique way. What is needed, therefore, is a ranking procedure that maps a preference relation to a maximally consistent ranking [24]. The difficulty of this problem depends on the concrete consistency criterion used, though many natural objectives (e.g., minimizing the number of object pairs whose ranks are in conflict with their pairwise preference) lead to NP-hard problems [2]. Fortunately, efficient techniques such as simple voting (known as the Borda count procedure in social choice theory) often deliver good approximations, sometimes even with provable guarantees [25], [26].

Another approach to learning ranking functions is to proceed from specific model assumptions, that is, assumptions about the structure of the preference relations. This approach is less generic than the previous ones, as it strongly depends on the concrete assumptions made. An example is the assumption that the target ranking of a set of objects described in terms of multiple attributes can be represented as a *lexicographic order* [27]–[29]. Another example is the assumption that the target ranking can be represented by a CP-net [30]. From a machine learning point of view, assumptions of the aforementioned type can be seen as an inductive bias restricting the hypothesis space. Provided the bias is correct, this is clearly an advantage, as it may simplify the learning problem considerably.

Yet another alternative is to resort to the idea of local estimation techniques as prominently represented, for example, by the nearest neighbor estimation principle [31], [32]: Considering the rankings observed in similar situations as representative, a ranking for the current situation is estimated on the basis of these "neighbored" rankings, typically using an averaging-like aggregation operator [18], [33]. This approach is in a sense orthogonal to the previous model-based one, as it is very flexible and typically comes with no specific model assumption (except the regularity assumption underlying the nearest neighbor inference principle).

B. The Choquet Integral in Machine Learning

Although the Choquet integral has been widely applied as an aggregation operator in multiple criteria decision making [34]–[36] and as a tool for preference elicitation [37], [38], it has been used much less in the field of machine learning so far. There are, however, a few notable exceptions.

Methods for binary classification based on the Choquet integral were developed in [39] and [40]. In [39], Grabisch and Roubens essentially employ the Choquet integral as a fusion operator in this context. For an instance $\boldsymbol{x} = (x_1, \ldots, x_n)$, let $\phi_i^{(j)}(\boldsymbol{x})$ express a measure of confidence (provided by feature i) that $\boldsymbol{x}$ belongs to class $j \in \{0, 1\}$. They define the global confidence for class j as an aggregation of these confidence degrees:

$$\phi_{\mu^{(j)}}(\boldsymbol{x}) \stackrel{\text{df}}{=} \mathcal{C}_{\mu^{(j)}}\left(\phi_1^{(j)}(\boldsymbol{x}), \ldots, \phi_n^{(j)}(\boldsymbol{x})\right),$$

where $\mathcal{C}_\mu$ denotes the discrete Choquet integral with respect to the fuzzy measure μ. Eventually, the class with the highest global confidence is predicted as an output. Here, the fuzzy measures $\mu^{(0)}$ and $\mu^{(1)}$ express the importance of the features and groups of features in the classification process. The $\phi_i^{(j)}$ are assumed to be derived by means of a conventional parametric or nonparametric probability density estimation method, subsequent to suitable normalization. The identification of the fusion operator is then reduced to the identification (or learning) of the fuzzy measures $\mu^{(0)}$ and $\mu^{(1)}$ with $2(2^n - 2)$ coefficients. To this end, Grabisch minimizes the empirical squared error loss

$$J = \sum_{\boldsymbol{x} \in T_0} \left(\phi_{\mu^{(0)}}(\boldsymbol{x}) - \phi_{\mu^{(1)}}(\boldsymbol{x}) - 1\right)^2 + \sum_{\boldsymbol{x} \in T_1} \left(\phi_{\mu^{(1)}}(\boldsymbol{x}) - \phi_{\mu^{(0)}}(\boldsymbol{x}) - 1\right)^2, \quad (1)$$

i.e., the sum of squared differences between predicted and given output values, using standard optimization routines (T_0 and T_1 denote, respectively, the set of observed negative and positive examples). Yan *et al.* [40] tackle a quite similar problem, albeit using another optimization criterion (which can be seen as a kind of relaxed class separability criterion). Besides, the authors define the Choquet integral based on a so-called signed nonadditive measure [41].

Apart from binary classification, the Choquet integral was also used in ordinal classification, a special type of multi-class classification in which the class labels are linearly ordered (e.g., a paper submitted for publication can be labeled as *accept*, *weak accept*, *weak reject*, or *reject*). Grabisch *et al.* [42], [43] consider input data of the following kind: a reference set of objects $A = \{1, \ldots, l\}$ and a set of criteria $X = \{1, \ldots, n\}$; a table of individual scores (performances) z_{ki} ($k \in A$, $i \in X$); a partial preorder $\geq_A$ on A (partial ranking of the objects on a global basis); a partial preorder $\geq_X$ on X (partial ranking of the criteria); a partial preorder $\geq_P$ on the set of pairs of criteria (partial ranking of interaction); the sign of interaction between selected pairs of criteria, reflecting synergy, independence, or redundancy. All this information can be translated into linear equalities or inequalities between the weights of the underlying fuzzy measure μ. This measure is then identified based on a constraint optimization problem, using as objective function a criterion that resembles very much the so-called margin principle in machine learning. The method itself, however, is more oriented toward decision making and less suitable for machine learning applications. In particular, it is not tolerant toward noise in the data and, in terms of complexity, does not scale well with the size of the data.

In [44], Beliakov and James develop a method for classifying journals in the field of pure mathematics, which are rated on an ordinal scale with categories A^+, A, B, and C. The classification is done on the basis of five criteria serving as input attributes, namely the number of citations per year, the impact factor, the immediacy index, the total number of articles published, and the cited half-life index (we shall use the same dataset in our experiments later on). As a loss function, the authors use the absolute difference between the predicted class and the target (i.e., the loss is $|i - j|$ if the ith class is predicted although the jth class would be correct).

Although our focus in this paper is on the use of the Choquet integral in supervised learning, it is worth mentioning that it can also be used in other settings. In the recent paper [45] by Beliakov *et al.*, the discrete Choquet integral is used for metric learning in semisupervised clustering. The authors investigate necessary and sufficient conditions for the discrete Choquet integral to define a metric, and, as a special case, obtain analogous conditions for ordered weighted averaging (OWA) operators. The corresponding metric learning problem is formulated as a linear programming problem.

III. The Discrete Choquet Integral

In this section, we give a brief introduction to the (discrete) Choquet integral, starting with a reminder of non-additive measures. In contrast to other fields, such as decision making and aggregation operators, the Choquet integral is not widely known in machine learning so far, which is the main reason to recall its definition in some detail. Readers familiar with the Choquet integral and related notions can safely skip this section.

A. Non-additive Measures

Let $X = \{x_1, \ldots, x_n\}$ be a finite set and μ a measure $2^X \to [0, 1]$. For each $A \subseteq X$, we interpret $\mu(A)$ as the *weight* or, say, the *importance* of the set of elements A. As an illustration, one may think of X as a set of criteria (binary features) relevant for a job, such as "speaking French" and "programming Java," and of $\mu(A)$ as the evaluation of a candidate satisfying criteria A (and not satisfying $X \setminus A$). The term "criterion" is indeed often used in the decision-making literature, where it suggests a monotone "the higher the better" influence.

A standard assumption on a measure $\mu(\cdot)$, which is, for example, at the core of probability theory, is additivity: $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \subseteq X$ such that $A \cap B = \emptyset$. Unfortunately, additive measures cannot model any kind of interaction between elements: extending a set of elements $A \subseteq X$ by a set of elements $B \subseteq X \setminus A$ always increases the weight $\mu(A)$ by the weight $\mu(B)$, regardless of A and B.

Suppose, for example, that the elements of two sets A and B are *complementary* in a certain sense. For instance, $A = \{\texttt{French}, \texttt{Spanish}\}$ and $B = \{\texttt{Java}\}$ could be seen as complementary, since both language skills and programming skills are important for the job. Formally, this can be expressed in terms of a positive interaction: $\mu(A \cup B) > \mu(A) + \mu(B)$.

In the extreme case, when language skills and programming skills are indeed essential, $\mu(A \cup B)$ can be high although $\mu(A) = \mu(B) = 0$ (suggesting that a candidate lacking either language or programming skills is completely unacceptable). Likewise, elements can interact in a negative way: if two sets A and B are partly *redundant* or *competitive*, then $\mu(A \cup B) < \mu(A) + \mu(B)$. For example, $B = \{\texttt{Java}\}$ and $C = \{\texttt{C}, \texttt{C\#}\}$ might be seen as redundant, since knowledge of one programming language does in principle suffice.

The above considerations motivate the use of non-additive measures, also called capacities or fuzzy measures, which are simply normalized and monotone:

$$\mu(\emptyset) = 0, \quad \mu(X) = 1, \quad \text{and}$$
$$\mu(A) \leq \mu(B) \text{ for all } A \subseteq B \subseteq X. \tag{2}$$

A useful representation of non-additive measures, which we shall explore later on for learning Choquet integrals, is in terms of the *Möbius transform*:

$$\mu(B) = \sum_{A \subseteq B} \boldsymbol{m}(A) \tag{3}$$

for all $B \subseteq X$, where the Möbius transform $\boldsymbol{m}_\mu$ of the measure μ is defined as follows:

$$\boldsymbol{m}_\mu(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \mu(B). \tag{4}$$

The value $\boldsymbol{m}_\mu(A)$ can be interpreted as the weight that is *exclusively* allocated to A, instead of being indirectly connected with A through the interaction with other subsets.

A measure μ is said to be k-order additive, or simply k-additive, if k is the smallest integer such that $\boldsymbol{m}(A) = 0$ for all $A \subseteq X$ with $|A| > k$. This property is interesting for several reasons. First, as can be seen from (3), it means that a measure μ can formally be specified by significantly fewer than 2^n values, which are needed in the general case. Second, k-additivity is also interesting from a semantic point of view: as will become clear in the following, this property simply means that there are no interaction effects between subsets $A, B \subseteq X$ whose cardinality exceeds k.

B. Importance of Criteria and Interaction

An additive (i.e., k-additive with $k = 1$) measure μ can be written as follows,

$$\mu(A) = \sum_{x_i \in A} \mu(\{x_i\}) = \sum_{x_i \in A} w_i$$

with $w_i = \mu(\{x_i\})$ the weight of x_i. Due to (2), these weights are nonnegative and such that $\sum_{i=1}^n w_i = 1$. In this case, there is obviously no interaction between the criteria x_i, i.e., the influence of x_i on the value of μ is independent of the presence or absence of any other x_j. Besides, the weight w_i is a natural quantification of the *importance* of x_i.

Measuring the importance of a criterion x_i becomes obviously more involved when μ is non-additive. Besides, one may then also be interested in a measure of *interaction* between the criteria, either pairwise or even of a higher order. In the literature, measures of that kind have been proposed, both for the importance of single as well as the interaction between several criteria.

Given a fuzzy measure μ on X, the *Shapley value* (or importance index) of x_i is defined as a kind of average increase in importance due to adding x_i to another subset $A \subset X$:

$$\varphi(x_i) = \sum_{A \subseteq X \setminus \{x_i\}} \frac{1}{n \binom{n-1}{|A|}} (\mu(A \cup \{x_i\}) - \mu(A)). \tag{5}$$

The Shapley value of μ is the vector $\varphi(\mu) = (\varphi(x_1), \ldots, \varphi(x_n))$. One can show that $0 \leq \varphi(x_i) \leq 1$ and $\sum_{i=1}^n \varphi(x_i) = 1$. Thus, $\varphi(x_i)$ is a measure of the *relative* importance of x_i. Obviously, $\varphi(x_i) = \mu(\{x_i\})$ if μ is additive.

The *interaction index* between criteria x_i and x_j, as proposed by Murofushi and Soneda [46], is defined as follows:

$$I(x_i, x_j) = \sum_{A \subseteq X \setminus \{x_i, x_j\}} \vartheta_A \cdot (\mu(A \cup \{x_i, x_j\}) - \mu(A \cup \{x_i\}) - \mu(A \cup \{x_j\}) + \mu(A)),$$

with

$$\vartheta_A = \frac{1}{(n-1)\binom{n-2}{|A|}}.$$

This index ranges between -1 and 1 and indicates a positive (negative) interaction between criteria x_i and x_j if $I_{i,j} > 0$ ($I_{i,j} < 0$). The interaction index can also be expressed in terms of the Möbius transform:

$$I(x_i, x_j) = \sum_{K \subseteq X \setminus \{x_i, x_j\}, |K|=k} \frac{1}{k+1} \boldsymbol{m}\Big(\{x_i, x_j\} \cup K\Big).$$

Furthermore, as proposed by Grabisch [47], the definition of interaction can be extended to more than two criteria, i.e., to subsets $T \subseteq X$:

$$I(T) = \sum_{k=0}^{n-|T|} \frac{1}{k+1} \sum_{K \subseteq X \setminus T, |K|=k} \boldsymbol{m}\Big(T \cup K\Big).$$

C. The Choquet Integral

So far, the criteria x_i were simply considered as binary features, which are either present or absent. Mathematically, $\mu(A)$ can thus also be seen as an *integral* of the indicator function of A, namely the function f_A given by $f_A(x) = 1$ if $x \in A$ and $= 0$ otherwise. Now, suppose that $f\colon X \to \mathbb{R}_+$ is any nonnegative function that assigns a *value* to each criterion x_i; for example, $f(x_i)$ might be the degree to which a candidate satisfies criterion x_i. An important question, then, is how to *aggregate* the evaluations of individual criteria, i.e., the values $f(x_i)$, into an overall evaluation, in which the criteria are properly weighted according to the measure μ. Mathematically, this overall evaluation can be considered as an integral $\mathcal{C}_\mu(f)$ of the function f with respect to the measure μ.

Indeed, if μ is an additive measure, the standard integral just corresponds to the *weighted mean*

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{n} w_i \cdot f(x_i) = \sum_{i=1}^{n} \mu(\{x_i\}) \cdot f(x_i), \qquad (6)$$

which is a natural aggregation operator in this case. A nontrivial question, however, is how to generalize (6) in the case where μ is non-additive.

This question, namely how to define the integral of a function with respect to a non-additive measure (not necessarily restricted to the discrete case), is answered in a satisfactory way by the Choquet integral, which has first been proposed for additive measures by Vitali [48] and later on for non-additive measures by Choquet [8]. The point of departure of the Choquet integral is an alternative representation of the "area" under the function f, which, in the additive case, is a natural interpretation of the integral. Roughly speaking, this representation decomposes the area into a "horizontal" instead of a "vertical" manner, thereby making it amenable to a straightforward extension to the non-additive case. More specifically, note that the weighted mean can be expressed as follows:

$$\begin{aligned}
&\sum_{i=1}^{n} f(x_i) \cdot \mu(\{x_i\}) \\
&= \sum_{i=1}^{n} \Big(f(x_{(i)}) - f(x_{(i-1)}) \Big) \cdot \Big(\mu(\{x_{(i)}\}) + \cdots + \mu(\{x_{(n)}\}) \Big) \\
&= \sum_{i=1}^{n} \Big(f(x_{(i)}) - f(x_{(i-1)}) \Big) \cdot \mu\Big(A_{(i)} \Big),
\end{aligned}$$

where $(\cdot)$ is a permutation of $\{1, \ldots, n\}$ such that $0 \leq f(x_{(1)}) \leq f(x_{(2)}) \leq \cdots \leq f(x_{(n)})$ (and $f(x_{(0)}) = 0$ by definition), and $A_{(i)} = \{x_{(i)}, \ldots, x_{(n)}\}$.

Now, the key difference between the left- and right-hand sides of the above expression is that, whereas the measure μ is only evaluated on single elements x_i on the left, it is evaluated on *subsets* of elements on the right. Thus, the right-hand side suggests an immediate extension to the case of non-additive measures, namely the Choquet integral, which, in the discrete case, is formally defined as follows:

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{n} (f(x_{(i)}) - f(x_{(i-1)})) \cdot \mu(A_{(i)}).$$

In terms of the Möbius transform of μ, the Choquet integral can also be expressed as follows:

$$\begin{aligned}
\mathcal{C}_\mu(f) &= \sum_{i=1}^{n} (f(x_{(i)}) - f(x_{(i-1)})) \cdot \mu(A_{(i)}) \\
&= \sum_{i=1}^{n} f(x_{(i)}) \cdot (\mu(A_{(i)}) - \mu(A_{(i+1)})) \\
&= \sum_{i=1}^{n} f(x_{(i)}) \sum_{R \subseteq \mathfrak{T}_{(i)}} \boldsymbol{m}(R) \\
&= \sum_{T \subseteq X} \boldsymbol{m}(T) \times \min_{x_i \in T} f(x_i) \qquad (7)
\end{aligned}$$

where $\mathfrak{T}_{(i)} = \{\mathfrak{S} \cup \{x_{(i)}\} \,|\, \mathfrak{S} \subset \{x_{(i+1)}, \ldots, x_{(n)}\}\}$. Note that expression (7) can also be written in terms of an inner product

$$\langle \boldsymbol{m}_\varphi, \varphi(f(\boldsymbol{x})) \rangle,$$

with the mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^{2^n - 1}$ defined as follows:

$$\begin{aligned}
\varphi(\boldsymbol{x}) &= \varphi(x_1, \ldots, x_n) \\
&= \Big(x_1, \ldots, x_n, \min\{x_1, x_2\}, \ldots, \min\{x_{n-1}, x_n\}, \\
&\qquad \min\{x_1, x_2, x_3\}, \ldots, \min\{x_1, \ldots, x_n\} \Big).
\end{aligned}$$

Moreover, $\boldsymbol{m}_\varphi$ denotes the vector $(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n, \boldsymbol{m}_{n+1}, \ldots, \boldsymbol{m}_{2^n - 1})$ of values of the Möbius transform in the order determined by $\varphi(\boldsymbol{x})$.

IV. Multipartite Ranking

As mentioned earlier, different types of ranking problems have recently been studied in the machine learning literature. Here, we are specifically interested in the so-called *multipartite ranking* problem [6].

In most ranking problems, the goal is to learn a *ranking function* that accepts a subset $\mathbf{O} \subset \mathcal{O}$ of objects as input, where $\mathcal{O}$ is a reference set of objects (e.g., the set of all books or movies). As output, the function produces a ranking (total order) $\succeq$ of the objects $\mathbf{O}$. Typically, a ranking function of that kind is implemented by means of a scoring function $U\colon \mathcal{O} \to \mathbb{R}$ so that

$$\boldsymbol{o} \succeq \boldsymbol{o}' \quad \Leftrightarrow \quad U(\boldsymbol{o}) \geq U(\boldsymbol{o}')$$

for all $\boldsymbol{o}, \boldsymbol{o}' \in \mathcal{O}$. Obviously, $U(\boldsymbol{o})$ can be considered as a kind of utility degree assigned to the object $\boldsymbol{o} \in \mathcal{O}$. Seen from this point of view, the goal in multipartite ranking is to learn a latent utility function on a reference set $\mathcal{O}$. In the following, we shall also refer to $U(\cdot)$ itself as a ranking function. Moreover, we assume that this function produces a strict order relation $\succ$, i.e., ties $U(\boldsymbol{o}) = U(\boldsymbol{o}')$ do not occur or are broken at random.

In order to induce a ranking function $U(\cdot)$, a learning algorithm (or "learner" for short) is provided with training information. In the case of multipartite ranking, the "ground truth" is supposed to be an ordinal categorization of objects. That is, each object $\boldsymbol{o} \in \mathcal{O}$ belongs to one of the classes in $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, where the classes are sorted such that $\lambda_1 < \lambda_2 < \cdots < \lambda_k$. Correspondingly, training data consist of a set of labeled objects $(\boldsymbol{o}_i, \ell_i) \in \mathcal{O} \times \mathcal{L}$, just like in ordinal regression.

The goal is to learn a ranking function $U(\cdot)$ that agrees well with the sorting of the classes in the sense that objects from higher classes are ranked higher than objects from lower classes. In [6], it was proposed to use the so-called C-index as a performance measure reflecting this goal in an adequate way:

$$C(U, \mathbf{O}) = \frac{\sum_{1 \leq i < j \leq k} \sum_{(\boldsymbol{o}, \boldsymbol{o}') \in \mathbf{O}_i \times \mathbf{O}_j} S(U(\boldsymbol{o}), U(\boldsymbol{o}'))}{\sum_{i<j} |\mathbf{O}_i| \cdot |\mathbf{O}_j|},$$

where $\mathbf{O}_i$ is the subset of objects $\boldsymbol{o} \in \mathbf{O}$ whose true class is λ_i, and

$$S(u,v) = \begin{cases} 1, & u < v \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

indicates whether or not a pair of objects has been ranked correctly. Thus, the C-index compares those object pairs $(\boldsymbol{o}, \boldsymbol{o}') \in \mathbf{O} \times \mathbf{O}$ where the class of $\boldsymbol{o}$ is lower than the class of $\boldsymbol{o}'$, and checks whether $U(\boldsymbol{o}) < U(\boldsymbol{o}')$, i.e., whether U correctly assigns a higher utility degree to $\boldsymbol{o}'$ than to $\boldsymbol{o}$. $C(U, \mathbf{O})$ is then simply given by the fraction of correct pairwise comparisons of this kind. In the case of two classes, C-index reduces to AUC (area under the ROC curve) that is widely used in binary classification.

V. Learning to Rank Using the Choquet Integral

The idea of our approach is to represent the latent utility function $U(\cdot)$ in terms of a Choquet integral. Assuming that objects $\boldsymbol{o} \in \mathcal{O}$ are represented as feature vectors

$$f_{\boldsymbol{o}} = (f_{\boldsymbol{o}}(x_1), \ldots, f_{\boldsymbol{o}}(x_n)),$$

where $f_{\boldsymbol{o}}(x_i)$ can be thought of as the evaluation of object $\boldsymbol{o}$ on the criterion x_i, this means that

$$U(\boldsymbol{o}) = \mathcal{C}_\mu(f_{\boldsymbol{o}}). \tag{9}$$

This approach appears to be interesting for a number of reasons, notably the following:

1) The representation (9) covers the commonly used linear utility functions as a special case.
2) Generalizing beyond the linear case, it is also able to capture more complex, nonlinear dependences and interactions between criteria.
3) The Choquet integral offers various means for explaining and understanding a utility function, including the importance value and the interaction index.
4) As opposed to many other models used in machine learning, the Choquet integral guarantees monotonicity in all criteria [49]. This is a reasonable property of a utility function which is often required in practice.

We assume training data to be available in the form of a set of objects $\boldsymbol{o}_i$ or, more specifically, the feature representation $f_{\boldsymbol{o}_i}$ of these objects, together with corresponding label information ℓ_i, $i = 1, \ldots, N$. From these data, a set D of pairwise preferences is constructed: $(\boldsymbol{o}_i, \boldsymbol{o}_j) \in D$, suggesting that $\boldsymbol{o}_i \succ \boldsymbol{o}_j$, if the training data contain $(\boldsymbol{o}_i, \ell_i)$ and $(\boldsymbol{o}_j, \ell_j)$ with $\ell_i > \ell_j$.

Following the idea of empirical risk minimization [9], we seek to induce a Choquet integral that minimizes the number of ranking errors (8) on the training data D. Since the Choquet integral is uniquely identified by the underlying measure μ on the set of criteria $X = \{x_1, \ldots, x_n\}$, this comes down to defining this measure in a most suitable way. In this regard, we make use of the representation (7) of μ in terms of its Möbius transform. Inspired by the maximum margin principle in kernel-based machine learning [9], we formulate the problem of learning μ as an optimization problem:

$$\max_{M, \xi_1, \ldots, \xi_N} \left\{ M - \frac{\gamma}{|D|} \sum_{(\boldsymbol{o}_s, \boldsymbol{o}_t) \in D} \xi_s + \xi_t \right\} \tag{10}$$

such that

$$\mathcal{C}_\mu(f_{\boldsymbol{o}_s}) - \mathcal{C}_\mu(f_{\boldsymbol{o}_t}) > M - \xi_s - \xi_t \quad \forall (\boldsymbol{o}_s, \boldsymbol{o}_t) \in D \tag{11}$$

$$\xi_s \geq 0 \qquad s \in \{1, \ldots, N\} \tag{12}$$

$$\sum_{T \subseteq X} \boldsymbol{m}(T) = 1 \tag{13}$$

$$\sum_{B \subseteq A} \boldsymbol{m}(B) \geq 0 \qquad \forall A \subseteq X \tag{14}$$

$$\sum_{L \subseteq A} \boldsymbol{m}(L) \leq \sum_{K \subseteq B} \boldsymbol{m}(K) \qquad \forall A \subset B \subseteq X \tag{15}$$

In this problem, M denotes the margin to be maximized, that is, the smallest difference between the utility degrees of two training objects $\boldsymbol{o}_s$ and $\boldsymbol{o}_t$ with $\boldsymbol{o}_s \succ \boldsymbol{o}_t$. More specifically, M is a *soft margin*: accounting for the fact that it will generally be impossible to satisfy all inequalities simultaneously, each object $\boldsymbol{o}_s$ is associated with a slack variable ξ_s. The slack variables are nonnegative, and a positive slack is penalized in proportion to its size. Finally, γ is a tradeoff parameter that controls the flexibility of the model; the higher the γ the stronger the slacks are punished.

A. Dealing With Constraints on the Fuzzy Measure

The constraints (13)–(15) formalize, respectively, the normalization, nonnegativity, and monotonicity of the Möbius transform. Obviously, the nonnegativity and monotonicity conditions are quite costly and produce as many as $3^n - 2^n$ constraints, since each subset of X is compared with all its subsets:

$$\sum_{i=1}^{n} \binom{n}{i} (2^i - 1) = \sum_{i=1}^{n} \binom{n}{i} 2^i - \sum_{i=1}^{n} \binom{n}{i} = 3^n - 2^n.$$

Fortunately, the last two constraints can be represented in a more compact way, exploiting a transitivity property:

$$\sum_{B \subseteq A \setminus \{x_i\}} \boldsymbol{m}(B \cup \{x_i\}) \geq 0 \qquad \forall A \subseteq X, x_i \in X.$$

This representation reduces the number of constraints to $n2^{n-1}$, which, despite still being large, is a significant reduction in comparison to the original formulation.

Another way of reducing complexity is to restrict the class of fuzzy measures to k-additive measures, that is, setting $\boldsymbol{m}(A) = 0$ for all $A \subseteq X$ with $|A| > k$. In fact, choosing $k \ll n$ is not only interesting from an optimization but also from a learning point of view: since the degree of additivity of μ offers a way to control the *capacity* of the underlying model class, selecting a proper k is crucial in order to guarantee the generalization performance of the learning algorithm. More specifically, the larger k is chosen, the more flexibly the Choquet integral can be fitted to the data. Thus, choosing k too large comes along with a danger of overfitting the data.

Algorithm 1 Cutting plane algorithm

1: Input: training set $S = \{(f_{i_1 j_1}, y_1), ..., (f_{i_K j_K}, y_K)\} \subset [-1,+1]^n \times \{-1,+1\}$, ζ, ϵ
2: $\mathcal{W} = \emptyset$
3: **repeat**
4: $(w, \xi) \leftarrow \arg\min_{w,\xi \geq 0} \frac{1}{2} w^\top w + \zeta\xi$
s.t.
5: $\forall(\overline{y}_1, \ldots, \overline{y}_K) \in \mathcal{W}$:
$w^\top \cdot \sum_{k=1}^K \Delta(y_k, \overline{y}_k) \left[\psi(f_{i_k j_k}, y_k) - \psi(f_{i_k j_k}, \overline{y}_k)\right] \geq \sum_{k=1}^K \Delta(y_k, \overline{y}_k) - K\xi$
6: $w \cdot (\phi(Z^*) - \phi(Z)) \leq 0$ for all $Z^* \in S_{l-1}^*$, $Z \in S_l^*$ s. t. $\sum_{k=1}^n z_k - z_k^* = 1$ $(1 \leq l \leq n)$
7: **for** $k = 1$ to K **do**
8: $\tilde{y}_k \leftarrow \arg\max_{\tilde{y} \in \{-1,+1\}} \Delta(y_k, \tilde{y}) \Big(1 - w^\top \cdot [\psi(f_{i_k j_k}, y_k) - \psi(f_{i_k j_k}, \tilde{y})]\Big)$
9: **end for**
10: $\mathcal{W} \leftarrow \mathcal{W} \cup \{(\tilde{y}_1, \ldots, \tilde{y}_K)\}$
11: **until**
$\sum_{k=1}^K \Delta(y_k, \tilde{y}_k) - w^\top \cdot \sum_{k=1}^K \Delta(y_k, \tilde{y}_k)[\psi(f_{i_k j_k}, y_k) - \psi(f_{i_k j_k}, \tilde{y}_k)] \leq K\xi + K\epsilon$
12: return(w,ξ)
13: $\boldsymbol{m}_\phi = \frac{1}{\sum_{k=1}^{2^n-1} w_k} \cdot w$

B. Dealing With Soft Margin Constraints

The number of soft margin constraints (11) and (12) may become quite large, too, as it scales quadratically with the size N of the training data. To cope with the complexity implied by these constraints, we refer to the idea underlying *cutting plane algorithms*, which originates from linear optimization theory and has already been used successfully in learning support vector machines for classification [50].

The key idea of cutting plane algorithms is to solve the problem by considering only a subset of all constraints, hoping that the rest will be satisfied, too. If this is not the case, then more constraints are added. More concretely, most algorithms start with an empty set of constraints and successively add the constraint that is maximally violated by the current solution.

The pseudocode of the cutting plane approach is shown in Algorithm 1. For each pair of objects $(\boldsymbol{o}_i, \boldsymbol{o}_j) \in D$ with $\boldsymbol{o}_i \succ \boldsymbol{o}_j$, we denote by f_{ij} the difference $f_{\boldsymbol{o}_i} - f_{\boldsymbol{o}_j}$, to which we assign the class label $+1$; conversely, $f_{ji} = f_{\boldsymbol{o}_j} - f_{\boldsymbol{o}_i}$ is assigned the class label -1. The function ψ is defined as $\psi(\boldsymbol{x}, y) = \phi(\boldsymbol{x})y$, where $y \in \{-1, +1\}$. Moreover, Δ is a loss function defined as

$$\Delta(y, \overline{y}) = \begin{cases} 0, & y = \overline{y} \\ L, & \text{otherwise}, \end{cases} \tag{16}$$

where L is a rescaling parameter that can be set to 1 without loss of generality. Roughly speaking, the lower the L, the stronger the slack ξ penalized; the importance of the slack in the objective function is controlled by the parameter ζ. Finally, $S_l^* = \{(z_1, \ldots, z_n) \in \{0,1\}^n \mid \sum_{i=1}^n z_i = l\}$, where n is the number of features and $l \in \{1, \ldots, n\}$.

As can be seen, the algorithm iteratively constructs a working set $\mathcal{W} = \mathcal{W}_1 \cup \cdots \cup \mathcal{W}_r$ of constraints, starting with an empty set $\mathcal{W} = \emptyset$. In each step, the algorithm finds the constraint that is mostly violated by the current solution w, ξ and adds it to the working set. Additionally, it guarantees the monotonicity constraints; the formulation in line 6 is provably equivalent to the inequality constraints (15). The algorithm terminates as soon as no further constraints are violated by more than the predefined precision ϵ.

VI. Experimental Results

A. Data

Preference learning data meeting the requirements of our setting are by far not as abundant as data for standard machine learning problems such as classification and regression. In particular, note that we require data in which the output is measured on an ordered categorical scale. Moreover, since the Choquet integral is a monotone aggregation operator, the data should be monotone in the sense that the output can be expected to increase with each input attribute.

In total, we managed to collect 15 datasets meeting these requirements, mainly from the UCI repository[1] and the WEKA machine learning framework [52]. These are all benchmark datasets commonly used for experimental purposes in machine learning. Besides, we collected a number of real-world datasets from other sources, namely data from an industrial polyester dyeing process [53] and data about the evaluation of mathematical journals [44]. Table I provides a summary of all datasets, which can be downloaded from our website.[2] In what follows, we give a brief description of these datasets.

- Color: The first dataset is color yield. It originates from an industrial polyester dyeing process that was also analyzed in [51]. Here, the output variable is the color yield, which has been measured as a function of three important factors of the production process: disperse dyes concentration, temperature, and time of dyeing. Corresponding experiments have been made for seven different colors, giving rise to seven datasets. Since the output variable is actually numeric, we turn it into an ordinal class variable

[1] http://archive.ics.uci.edu/ml/
[2] http://www.uni-marburg.de/fb12/kebi/research/

TABLE I
DATASETS AND THEIR PROPERTIES

data set	#instances	#attributes	#classes	source
Color (CLR) 1–7	120	3	3	[51]
Scientific Journals (SCJ)	172	5	4	[44]
CPU	209	6	2	UCI
Auto MPG	398	8	6	UCI
Employee Selection (ESL)	488	4	9	WEKA
Mamographic (MMG)	830	5	2	UCI
Lecturers Evaluation (LEV)	1000	4	5	WEKA
Concrete Compressive Strength (CCS)	1030	8	6	UCI
Car Evaluation (CEV)	1728	6	4	UCI

with three values; to this end, two thresholds are defined in such a way that the class distribution is uniform.

- Scientific Journals: This dataset is comprised of journals in the field of pure mathematics, which are rated on a scale with categories A^+, A, B, and C [44]. Each journal is, moreover, scored in terms of five criteria serving as input attributes, namely, cites (the total number of citations per year), the impact factor (average number of citations per article within two years after publication), the immediacy index (cites to articles in current calendar year divided by the number of articles published in that year), articles (the total number of articles published), and cited half-life (median age of articles cited).
- CPU: This is a standard benchmark dataset from the UCI repository. It contains nine attributes, three of which were removed since they are obviously of no predictive value (vendor name, model name, ERP).
- Auto MPG: This dataset contains eight attributes and one output. The attributes are cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name. The last attribute (car name) was removed because it has no predictive value. The output is fuel consumption in miles per gallon (MPG). In order to obtain an ordinal class structure, the MPG value was discretized into six consecutive intervals.
- Employee Selection: This dataset contains profiles of applicants for certain industrial jobs. The values of the four input attributes were determined by psychologists based upon psychometric test results and interviews with the candidates. The output is an overall score on an ordinal scale between 1 and 9, corresponding to the degree of suitability for each candidate to this type of job.
- Mammographic: This dataset is about breast cancer screening by mammography. The goal is to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes (mass shape, mass margin, and density) and the patient's age.
- Lecturers' Evaluation: This dataset contains examples of anonymous lecturer evaluations, taken at the end of MBA courses. Students were asked to score their lecturers according to four attributes such as oral skills and contribution to their professional/general knowledge. The output was a total evaluation of each lecturer's performance, measured on an ordinal scale from 0 to 4.
- Concrete Compressive Strength: This dataset comprises eight quantitative input variables, namely the following features of concrete: cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age. The output set is the concrete compressive strength measured in megapascal. We turned it into an ordinal attribute using equiwidth binning with six bins.
- Car Evaluation: This dataset contains six attributes describing a car: buying price, price of the maintenance, number of doors, capacity in terms of persons to carry, the size of luggage boot, and estimated safety of the car. The output is the overall evaluation of the car: unacceptable, acceptable, good, and very good.

B. *Comparison With Linear and Polynomial Kernel Methods*

We compared our approach (subsequently referred to as CI) with kernel-based methods for ranking, using the spider implementation[3] of the RankSVM approach with a linear and a polynomial kernel [54]. A comparison with this class of methods is interesting for several reasons. First, kernel-based methods belong to the state of the art in the field of learning to rank. Second, they make use of the same type of learning algorithm (large margin maximization). Third, the use of a polynomial kernel leads to a model that bears some resemblance with a Choquet integral. In fact, using a polynomial kernel of degree d on the original feature representation of objects, i.e., a kernel of the form

$$K(\boldsymbol{o}, \boldsymbol{o}') = (\langle f_{\boldsymbol{o}}, f_{\boldsymbol{o}'} \rangle + \lambda)^d \,, \tag{17}$$

essentially comes down to fitting a linear model in an expanded feature space, in which the original features $f(x_1), \ldots, f(x_n)$ are complemented by all monomials of order $\leq d$. Thus, a polynomial kernel of degree d captures the same level of interactions between criteria as a Choquet integral on a k-additive fuzzy measure, when $k = d$. Note, however, that it does not guarantee monotonicity in the input attributes.

Moreover, we compared with the weighted mean (WM), which does indeed assure monotonicity, but which is not able to capture any interaction between variables. This model was implemented as a special case of our method, namely the case of the Choquet integral based on a 1-additive measure.

[3]http://people.kyb.tuebingen.mpg.de/spider/

TABLE II
PERFORMANCE IN TERMS OF THE AVERAGE C-INDEX ± STANDARD DEVIATION

data set	WM	PL d=1	PL d=2	PL d=3	CI
CLR 1	.9663±.0148(4)	.9506±.0155(5)	.9674±.0129(3)	.9700±.0141(2)	.9828±.0090(1)
CLR 2	.8740±.0293(4)	.8601±.0294(5)	.8876±.0200(3)	.9341±.0244(2)	.9804±.0128(1)
CLR 3	.9343±.0204(4)	.9268±.0219(5)	.9375±.0156(3)	.9633±.0143(2)	.9878±.0150(1)
CLR 4	.9357±.0171(4)	.9228±.0247(5)	.9431±.0189(3)	.9659±.0166(2)	.9915±.0056(1)
CLR 5	.9518±.0194(3)	.9485±.0179(5)	.9565±.0142(2)	.9516±.0171(4)	.9682±.0140(1)
CLR 6	.9046±.0202(4)	.8923±.0205(5)	.9127±.0201(3)	.9460±.0191(2)	.9825±.0121(1)
CLR 7	.8880±.0312(4)	.8797±.0256(5)	.8892±.0219(3)	.9258±.0237(2)	.9688±.0167(1)
SCJ	.8168±.0105(4)	.8098±.0112(5)	.8270±.0241(3)	.8313±.0109(2)	.8450±.0201(1)
CPU	.9965±.0027(3)	.9950±.0093(5)	.9978±.0012(2)	.9955±.0005(4)	.9986±.0014(1)
MPG	.8887±.0176(4)	.8850±.0143(5)	.8912±.0078(3)	.8967±.0093(2)	.9060±.0111(1)
ESL	.9497±.0162(2)	.9559±.0071(1)	.9465±.0104(4)	.9491±.0126(3)	.9424±.0098(5)
MMG	.8961±.0230(2)	.8536±.0168(4)	.8714±.0181(3)	.7813±.0350(5)	.9015±.0210(1)
LEV	.8710±.0289(2)	.8620±.0320(3)	.8713±.0250(1)	.8527±.0300(5)	.8610±.0320(4)
CCS	.8650±.0068(4)	.8586±.0102(5)	.8862±.0184(3)	.8962±.0203(2)	.9050±.0038(1)
CEV	.8981±.0066(4)	.8804±.0076(5)	.9118±.0059(3)	.9585±.0090(2)	.9771±.0039(1)
average rank	3.47	4.53	2.8	2.73	1.47

Additionally, the rank of each method it shown in brackets.

TABLE III
WIN STATISTICS (NUMBER OF DATASETS ON WHICH THE FIRST METHOD WAS BETTER THAN THE SECOND ONE)

	WM	PL d=1	PL d=2	PL d=3	CI
WM	–	14	2	5	2
PL d=1	1	–	1	3	2
PL d=2	13	14	–	4	2
PL d=3	10	12	11	–	1
CI	13	13	13	14	–

C. Experimental Setup

In order to assure commensurability [55], all features were normalized to the range 0 and 1 before learning, thereby turning them into a criterion, i.e., a "the higher the better" attribute. To this end, the transformation $f_{\boldsymbol{o}}(x_i) = (x_i - m_i)/(M_i - m_i)$ was used, where m_i and M_i are, respectively, the lower and upper bounds for x_i (estimated from the data); if the influence of x_i is actually negative, the mapping $f_{\boldsymbol{o}}(x_i) = (M_i - x_i)/(M_i - m_i)$ is used instead.

We randomly split the data into two parts, one half for training and one half for testing. In order to find the tradeoff parameter γ, we conducted a fivefold cross validation on the training data, selecting an optimal γ from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. The model induced from the training data is then evaluated on the test data, measuring performance in terms of the C-index. This procedure was repeated 100 times, and the results were averaged.

D. Results

An overview of the results is given in Table II. Moreover, Table III provides a summary in terms of win–loss statistics, showing, for each pair of methods, on how many datasets the first one is better than the second one. The overall picture conveyed by the results is clearly in favor of our method. In fact, the superiority of CI can also be corroborated statistically, noting that 12 wins are enough to reject the null hypothesis of equal performance according to a simple (two-tailed) sign test at the 5% significance level. Statistically, the results, thus, suggest that CI is signifiantly better than the competitor methods.

Table IV shows results for our method when restricting the Choquet integral to k-additive measures, for different values of k. A restriction of this kind is interesting for several reasons. First, since less parameters need to be estimated, it reduces complexity and, therefore, increases the efficiency of our learning algorithm. Second, as already explained earlier, a restriction to k-additive measures is also interesting from a learning (induction) point of view, as it allows for controlling the capacity of the underlying hypothesis space: the larger the value of k, the richer the model space. In other words, k can be used to control the flexibility (nonlinearity) of the model class. If k is too small, the model is not able to fit the data sufficiently well. On the other hand, if k is too large, there is a danger of overfitting the data, which may lead to poor generalization. Ideally, k is chosen so as to avoid both problems, namely under- and overfitting the data. As can be seen in Table IV, the ideal value does indeed depend on the dataset and is often smaller than the largest possible value (namely $k = \#$attributes). The question of how to determine an optimal value of k in an efficient way (i.e., without simply trying all alternatives) is an important topic of future work.

As one of the key features of our approach, we already mentioned the aspect of interpretability. In particular, the Choquet integral (or, more specifically, the underlying fuzzy measure) provides natural measures of the importance of individual and the interaction between pairs (or even groups) of attributes. As an illustration, Fig. 1 visualizes the (pairwise) interaction between attributes for the car evaluation data, for which CI performs significantly better than WM. Recall that, in this dataset, the evaluation of a car (output attribute) depends on a number of criteria, namely (a) buying price, (b) price of the maintenance, (c) number of doors, (d) capacity in terms of persons to carry, (e) size of luggage boot, and (f) safety of the car. These criteria form a natural hierarchy: (a) and (b) form a subgroup PRICE, whereas the other properties are of TECHNICAL nature and can be further decomposed into COMFORT (c)–(e) and safety (f). Interestingly, the interaction in our model nicely agrees with this hierarchy: interaction within each subgroup tends to be smaller (as can be seen from the darker colors) than the interaction between criteria from different subgroups, suggesting a kind of redundancy in the former and complementarity in the latter case.

In addition, Fig. 2 visualizes the interaction between the three attributes in the color yield datasets, namely for CLR-1 and

TABLE IV
C-INDEX FOR RESTRICTION TO k-ADDITIVE MEASURES (BEST RESULT HIGHLIGHTED IN BOLD)

data set	CI k=2	CI k=3	CI k=4	CI k=5	CI k=6
CLR 1	.9825±.0103	**.9828**±.0090	———	———	———
CLR 2	.9803±.0095	**.9804**±.0128	———	———	———
CLR 3	.9874±.0101	**.9878**±.0150	———	———	———
CLR 4	**.9921**±.0063	.9915±.0056	———	———	———
CLR 5	**.9683**±.0141	.9682±.0140	———	———	———
CLR 6	**.9867**±.0120	.9825±.0121	———	———	———
CLR 7	.9665±.0162	**.9688**±.0167	———	———	———
SCJ	**.8459**±.0245	.8447±.0201	.8458±.0191	.8450±.0201	———
CPU	.9986±.0009	.9985±.0016	**.9988**±.0010	.9984±.0011	.9986±.0014
MPG	.9038±.0136	.9044±.0111	.9049±.0104	.9058±.0093	**.9060**±.0111
ESL	**.9486**±.0172	.9457±.0176	.9424±.0098	———	———
MMG	.8941±.0290	.8960±.0217	.8964±.0172	**.9015**±.0210	———
LEV	**.8667**±.0216	.8635±.0241	.8610±.0320	———	———
CCS	**.9149**±.0131	.9123±.0155	.9120±.0184	.9104±.0104	.9050±.0038
CEV	.9411±.0082	.9660±.0066	.9670±.0053	.9678±.0041	**.9771**±.0039

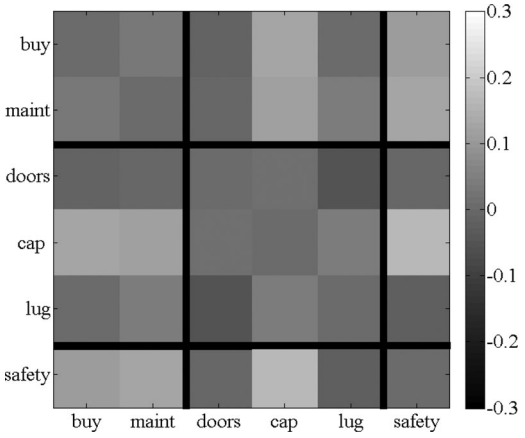

Fig. 1. Visualization of the interaction index for the car evaluation data (numerical values are shown in terms of level of gray; values on the diagonal are set to 0). Groups of related criteria are indicated by the black lines.

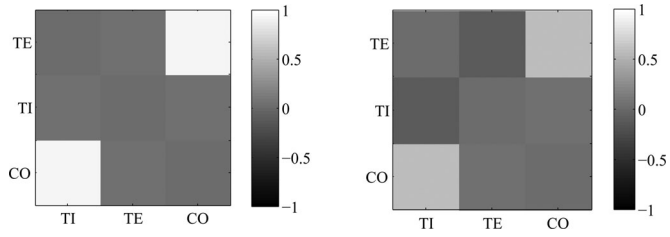

Fig. 2. Visualization of the interaction index for datasets CLR-1 (left) and CLR-7 (right). The reduction of prediction error is about twice as much as for CLR-7. For the ease of representation, the values on the diagonal are set to 0.

CLR-7. The interaction is not very strong in the case of CLR-1, but more pronounced for CLR-7. This is in agreement with the improvement achieved by CI in comparison with a simple linear model (WM), which is relatively small in the former but much higher in the latter case. In fact, it is plausible that a complex, nonlinear model like CI is not needed unless the attributes are strongly interacting. Or, stated differently, if there is (almost) no interaction between the attributes, a simple linear model will generally be enough.

VII. SUMMARY AND CONCLUSIONS

In this paper, we have advocated the use of the discrete Choquet integral in the context of preference learning. More specifically, we have used the Choquet integral for representing a latent utility function in multipartite ranking, a specific type of preference learning problem. This idea is motivated by several appealing properties offered by the Choquet integral, making it quite attractive from a preference learning point of view. This includes its ability to capture dependences between criteria (attributes) and to obey natural monotonicity conditions, as well as its interpretability. In fact, in preference learning, one is often not only interested in the prediction of preferences. Instead, it may also be important to get an explanation of a prediction, i.e., a reason for *why* an alternative A is (presumably) preferred to another alternative B. The measures of importance and interaction between attributes, which can directly be derived from the fuzzy measure underlying the Choquet integral, provide extremely valuable information in this regard.

The proper specification of this measure, i.e., the adaptation of the fuzzy measure to the data at hand, is the main challenge from a machine learning point of view. We formalized this problem as a (soft) margin maximization problem and solved it by means of a cutting plane algorithm. Our algorithm was compared with state-of-the-art ranking methods on a number of benchmark datasets. The results of these experiments are very promising and clearly in favor of our approach.

Needless to say, the method proposed in this paper can be refined in several directions. Especially interesting in this regard is the idea of restricting the model class to k-additive measures, connected with the use of k as a kind of regularization parameter (cf., Section VI). Moreover, going beyond the specific problem of multipartite ranking, one may of course also think of applying the Choquet integral to other types of preference learning problems. Indeed, being convinced of the high potential of this idea, we consider this paper as a first step toward establishing the Choquet integral as an important mathematical tool of preference learning, and a precursor for research along similar lines.

ACKNOWLEDGMENTS

The authors would like to thank M. Nasiri, University of Siegen, Germany, and G. Beliakov, Deakin University, Australia, for providing them the color yield data and the journal data, respectively. This work was supported by the German Research Foundation (DFG).

REFERENCES

[1] J. Fürnkranz and E. Hüllermeier, Eds., *Preference Learning*. New York: Springer-Verlag, 2010.

[2] W. Cohen, R. Schapire, and Y. Singer, "Learning to order things," *J. Artif. Intell. Res.*, vol. 10, no. 1, pp. 243–270, 1999.

[3] S. Har-Peled, D. Roth, and D. Zimak, "Constraint classification for multiclass classification and ranking," in *Advances in Neural Information Processing Systems,* S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15, Cambridge, MA: MIT Press, 2003, pp. 785–792.

[4] J. Fürnkranz and E. Hüllermeier, "Pairwise preference learning and ranking," in *Proc. Eur. Conf. Machine Learn. Principles Practice Knowledge Discovery Databases*, 2003, pp. 145–156.

[5] O. Dekel, C. Manning, and Y. Singer, "Log-linear models for label ranking," in *Advances in Neural Information Processing Systems,* S. Thrun, L. Saul, and B. Schölkopf, Eds., vol. 16, Cambridge, MA: MIT Press, 2003.

[6] J. Fürnkranz, E. Hüllermeier, and S. Vanderlooy, "Binary decomposition methods for multipartite ranking," in *Proc. Eur. Conf. Mach. Learn. Principles Practice Knowledge Discovery Databases*, 2009, pp. 359–374.

[7] G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, Eds., *Predicting Structured Data*. Cambridge, MA: MIT Press, 2007.

[8] G. Choquet, "Theory of capacities," *Ann. l'institut Four.*, vol. 5, pp. 131–295, 1954.

[9] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2001.

[10] U. Chajewska, L. Getoor, J. Norman, and Y. Shahar, "Utility elicitation as a classification problem," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 79–88.

[11] P. Haddawy, V. Ha, A. Restificar, B. Geisler, and J. Miyamoto, "Preference elicitation via theory refinement," *J. Mach. Learn. Res.*, vol. 4, pp. 317–337, 2003.

[12] P. Viappiani, P. Pu, and B. Faltings, "Preference-based search with adaptive recommendations," *AI Commun.—Recommend. Syst.*, vol. 21, no. 2–3, pp. 155–175, 2008.

[13] M. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artif. Intell. Rev.*, vol. 13, no. 5–6, pp. 393–408, 1999.

[14] P. Perny and J. Zucker, "Preference-based search and machine learning for collaborative filtering: The "film-conseil" movie recommender system," *Inf. Interact. Intell.*, vol. 1, no. 1, pp. 9–48, 2003.

[15] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, 2002, pp. 133–142.

[16] F. Radlinski and T. Joachims, "Query chains: Learning to rank from implicit feedback," in *Proc. 11th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, 2005, pp. 239–248.

[17] T. Kamishima, H. Kazawa, and S. Akaho, "A survey and empirical comparison of object ranking methods," in *Preference Learning*, J. Fürnkranz and E. Hüllermeier, Eds. New York: Springer-Verlag, 2010, pp. 181–202.

[18] K. Brinker and E. Hüllermeier, "Case-based label ranking," in *Proc. Eur. Conf. Mach. Learning Principles Practice Knowledge Discovery Databases*, 2006, pp. 566–573.

[19] S. Har-Peled, D. Roth, and D. Zimak, "Constraint classification: A new approach to multiclass classification," in *Proc. 13th Int. Conf. Algorithmic Learning Theory*, 2002, pp. 365–379.

[20] W. Cheng, J. Hühn, and E. Hüllermeier, "Decision tree and instance-based learning for label ranking," in *Proc. 26th Int. Conf. Mach. Learning*, 2009, pp. 161–168.

[21] S. Vembu and T. Gärtner, "Label ranking: A survey," in *Preference Learning*, J. Fürnkranz and E. Hüllermeier, Eds. New York: Springer-Verlag, 2010, pp. 45–64.

[22] J. Quevedo, E. Montanes, O. Luaces, and J. del Coz, "Adapting decision DAGs for multipartite ranking," in *Proc. Eur. Conf. Mach. Learning Principles Practice Knowledge Discovery Databases*, 2010, pp. 115–130.

[23] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker, "Label ranking by learning pairwise preferences," *Artif. Intell.*, vol. 172, pp. 1897–1917, 2008.

[24] E. Hüllermeier and J. Fürnkranz, "Comparison of ranking procedures in pairwise preference learning," presented at the 10th Int. Conf. Inform. Process. Management Uncertainty Knowledge-Based Syst., Perugia, Italy, 2004.

[25] D. Coppersmith, L. Fleischer, and A. Rudra, "Ordering by weighted number of wins gives a good ranking for weighted tournaments," *ACM Trans. Algorithms*, vol. 6, no. 3, pp. 55:1–55:13, 2010.

[26] E. Hüllermeier and J. Fürnkranz, "On predictive accuracy and risk minimization in pairwise label ranking," *J. Comput. Syst. Sci.*, vol. 76, no. 1, pp. 49–62, 2010.

[27] P. Flach and T. Matsubari, "A simple lexicographic ranker and probability estimator," in *Proc. Eur. Conf. Mach. Learning Principles Practice Knowledge Discovery Databases*, 2007, pp. 575–582.

[28] F. Yaman, T. Walsh, M. Littman, and M. des Jardins, "Democratic approximation of lexicographic preference models," in *Proc. 25th Int. Conf. Mach. Learning*, 2008, pp. 1200–1207.

[29] R. Booth, Y. Chevaleyre, J. Lang, J. Mengin, and C. Sombattheera, "Learning conditionally lexicographic preference relations," in *Proc. 19th Eur. Conf. Artif. Intell.*, 2010, pp. 269–274.

[30] Y. Chevaleyre, F. Koriche, J. Lang, J. Mengin, and B. Zanuttini, "Learning ordinal preferences on multiattribute domains: The case of CP-nets," in *Preference Learning*, J. Furnkranz and E. Hüllermeier, Eds. New York: Springer-Verlag, 2010, pp. 273–296.

[31] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE *Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[32] E. Hüllermeier, "Similarity-based inference as evidential reasoning," *Int. J. Approx. Reason.*, vol. 26, pp. 67–100, 2001.

[33] W. Cheng, K. Dembczyński, and E. Hüllermeier, "Label ranking based on the Placket–Luce model," in *Proc. 27th Int. Conf. Mach. Learning*, 2010, pp. 215–222.

[34] M. Grabisch, "Fuzzy integral for classification and feature extraction," in *Fuzzy Measures and Integrals: Theory and Applications*, M. Grabisch, T. Murofushi, and M. Sugeno, Eds. Heidelberg, Germany: Physica, 2000, pp. 415–434.

[35] M. Grabisch, "Fuzzy integral in multicriteria decision making," *Fuzzy Sets Syst.*, vol. 69, no. 3, pp. 279–298, 1995.

[36] V. Torra, "Learning aggregation operators for preference modeling," in *Preference Learning*, J. Fürnkranz and E. Hüllermeier, Eds. New York: Springer-Verlag, 2011, pp. 317–333.

[37] S. Angilella, S. Greco, and B. Matarazzo, "Non-additive robust ordinal regression with Choquet integral, bipolar and level dependent Choquet integrals," in *Proc. Joint Int. Fuzzy Syst. Assoc. World Congr. Eur. Soc. Fuzzy Logic Technol. Conf.*, 2009, pp. 1194–1199.

[38] S. Angilella, S. Greco, and B. Matarazzo, "Non-additive robust ordinal regression: A multiple criteria decision model based on the Choquet integral," *Eur. J. Operat. Res.*, vol. 201, pp. 277–288, 2010.

[39] M. Grabisch and M. Roubens, "Application of the Choquet integral in multicriteria decision making," in *Fuzzy Measures and Integrals: Theory and Applications*, M. Grabisch, T. Murofushi, and M. Sugeno, Eds. Heidelberg, Germany: Physica, 2000, pp. 348–374.

[40] N. Yan, Z. Wang, and Z. Chen, "Classification with Choquet integral with respect to signed non-additive measure," in *Proc. 7th IEEE Int. Conf. Data Mining Workshops*, 2007, pp. 283–288.

[41] T. Murofushi, M. Sugeno, and M. Machida, "Non-monotonic fuzzy measures and the Choquet integral," *Fuzzy Sets Syst.*, vol. 64, no. 1, pp. 73–86, 1994.

[42] M. Grabisch, T. Murofushi, and M. Sugeno, Eds., Fuzzy Measures and Integrals: Theory and Applications. Heidelberg, Germany: Physica, 2000.

[43] M. Grabisch, I. Kojadinovic, and P. Meyer, "A review of methods for capacity identification in Choquet integral based multi-attribute utility theory: Applications of the Kappalab R package," *Eur. J. Operat. Res.*, vol. 186, no. 2, pp. 766–785, 2008.

[44] G. Beliakov and S. James, "Citation-based journal ranks: The use of fuzzy measures," *Fuzzy Sets Syst.*, vol. 167, no. 1, pp. 101–119, 2011.

[45] G. Beliakov, S. James, and G. Li, "Learning Choquet-integral-based metrics for semisupervised clustering," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 3, pp. 562–574, Jun. 2011.

[46] T. Murofushi and S. Soneda, "Techniques for reading fuzzy measures (III): Interaction index," in *Proc. 9th Fuzzy Syst. Symp.*, 1993, pp. 693–696.
[47] M. Grabisch, "k-order additive discrete fuzzy measures and their representation," *Fuzzy Sets Syst.*, vol. 92, no. 2, pp. 167–189, 1997.
[48] G. Vitali, "Sulla definizione di integrale delle funzioni di una variabile," *Ann. Matematica Pura Appl.*, vol. 2, no. 1, pp. 111–121, 1925.
[49] A. F. Tehrani, W. Cheng, K. Dembczynski, and E. Hüllermeier, "Learning monotone nonlinear models using the Choquet integral," in *Proc. Eur. Conf. Mach. Learning Principles Practice Knowledge Discovery Databases*, 2011, pp. 414–429.
[50] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural SVM," *Mach. Learn.*, vol. 77, no. 1, pp. 27–59, 2009.
[51] M. Nasiri and S. Berlik, "Modeling of polyester dyeing using an evolutionary fuzzy system," in *Proc. Joint Int. Fuzzy Syst. Assoc. World Congr. Eur. Soc. Fuzzy Logic Technol. Conf.*, 2009, pp. 1246–1251.
[52] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorat. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[53] M. Nasiri, "Fuzzy regression modeling of colour yield in dyeing polyester with disperse dyes" Master's thesis, Textile Eng. Dept., Isfahan Univ. Technol., Isfahan, Iran, 2003.
[54] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 115–132.
[55] F. Modave and M. Grabisch, "Preference representation by a Choquet integral: Commensurability hypothesis," in *Proc. 7th Int. Conf. Inf. Proces. Manage. Uncertainty Knowledge-Based Syst.*, 1998, pp. 164–171.

Ali Fallah Tehrani received the B.Sc. degree in pure mathematics from Tehran University, Tehran, Iran, and the M.S. degree in mathematical logic and theoretical informatics from Tarbiat Modares University, Tehran. He is currently working toward the Ph.D. degree with the Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany, under the supervision of Prof. E. Hüllermeier.

He is a Research Staff Member with the Department of Mathematics and Computer Science, University of Marburg. His research interests include monotone learning, preference learning, fuzzy integrals, and data mining.

Weiwei Cheng received the Bachelor's degree in computer science and business administration from Zhengzhou University, Zhengzhou, China, and the Master's degree in data and knowledge engineering from the University of Magdeburg, Magdeburg, Germany. He is currently working toward the Ph.D. degree with the Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany, under the supervision of Prof. E. Hüllermeier.

He is a Research Staff Member with the Department of Mathematics and Computer Science, University of Marburg. His research interests include multiple areas in machine learning, such as supervised ranking, preference learning, and multilabel classification.

Mr. Cheng has received several research awards, including the Best Student Paper Award at the 19th European Conference on Machine Learning. He also received the 2009 Chinese Government Award for Outstanding Self-Financed Students Abroad.

Eyke Hüllermeier received the M.Sc. degrees in mathematics and business computing, the Ph.D. degree in computer science, and the Habilitation degree, all from the University of Paderborn, Paderborn, Germany.

He is a Full Professor with the Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany. His research interests include machine learning and data mining, fuzzy set theory, uncertainty and approximate reasoning, and applications in bioinformatics. He has published numerous research papers on these topics in top-grade journals and major international conferences.

Dr. Hüllermeier is a member of the IEEE Computational Intelligence Society and a board member of the European Society for Fuzzy Logic and Technology (EUSFLAT). He is a Coeditor-in-Chief of *Fuzzy Sets and Systems* and is on the editorial board of several other journals. Moreover, he is a Coordinator of the EUSFLAT working group on Learning and Data Mining and the Head of the IEEE Computational Intelligence Society Task Force on Machine Learning.