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

Ali Fallah Tehrani, Weiwei Cheng and Eyke Hüllermeier

Dept. of Mathematics and Computer Science
University of Marburg, Germany
{fallah, cheng, eyke}@mathematik.uni-marburg.de

Abstract

In this paper, we propose a novel method for two types of ranking problems that have recently been introduced in the context of preference learning, an emerging subfield of machine learning. In the literature, these problems are referred to, respectively, as *object ranking* and *multipartite ranking*. In both cases, the task is to learn a ranking model that accepts as input a subset of alternatives, with each alternative typically represented in terms of a feature vector, and produces a ranking of these alternatives as output. Our approach is based on the idea of using the (discrete) Choquet integral as an underlying model for representing rankings. Being an established aggregation function in multiple criteria decision making and information fusion, the Choquet integral offers a number of interesting properties that render it attractive from a machine learning perspective, too. The learning problem itself, which comes down to properly specifying the fuzzy measure on which the Choquet integral is defined, is formalized as a margin maximization problem. For testing the performance of our method, we apply it to a real problem, namely the ranking of scientific journals.

1 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 revealing 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 novel methods for two types of ranking problems, using the (discrete) Choquet integral [8] as an underlying model for representing rankings. 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 redundancy, complementarity and interaction between different features. Moreover, it obeys certain 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].

We develop learning algorithms for two types of problems that have been referred to, respectively, as *object ranking* and *multipartite ranking* in the literature [2, 6]. In both cases, the task is to learn a ranking model that accepts as input an arbitrary set of alternatives, with each alternative typically represented in terms of a feature vector, and produces a ranking of these alternatives as output. The main difference concerns the training information, which is given in the form of *absolute* judgments in multipartite ranking and *relative* judgments in object ranking. More specifically, it consists of a set of evaluated alternatives in the former case, rated in terms of preference degrees on an ordinal scale (such as bad, good, very good), and of a set of pairwise comparisons between alternatives in the second case (suggesting that one alternative is preferred to another one).

For testing the performance of our methods, we apply them to a real problem, namely the ranking of scientific journals based on various properties and indicators, such as impact factor. A corresponding data set will not only be used to compare our methods with existing approaches in terms of predictive performance but also to highlight the advantages of the Choquet integral from a modeling and knowledge representation point of view.

The rest of this paper is organized as follows. In the next section, we give a brief overview of related work. In Section 3, we recall the basic definition of the Choquet integral and related notions. The ranking problems we are dealing with are explained in Section 4, and our approach for tackling them is introduced in Section 5. Finally, some first experimental results are presented in Section 6.

2 Related Work

Although the Choquet integral has been widely applied as an aggregation operator in multiple criteria decision making [10–12], it has been used much less in the field of machine learning so far. There are, however, a few notable exceptions.

First, the problem of extracting a Choquet integral (or, more precisely, the non-additive measure on which it is defined) in a data-driven way has been addressed in the literature. Essentially, this is a parameter identification problem, which is commonly formalized as a constraint optimization problem, for example using the sum of squared errors as an objective function [13, 14]. To this end, a heuristic, gadient-based method called HLMS (Heuristic Least Mean Squares) was introduced in [15], while [16] proposed an alternative approach based on the use of quadratic forms. Besides, genetic algorithms have been used as a tool for parameter optimization [17]. Some mathematical results regarding this optimization problem can be found in [18, 19].

Second, the Choquet integral has been used in a few works for learning classification models. Recently, for example, it has been used for ordinal classification [20,21]. In [22], the problem of learning an optimal classification function is cast in the setting of margin-maximization. Although the learning problem is different, this approach is especially relevant for us, since we shall employ quite similar techniques (cf. Section 5).

3 The Discrete Choquet Integral

In this section, we recall the basic definition of the Choquet integral and related notions. The first definition of the Choquet integral for additive measures is due to Vitali [23]. For the general case of a capacity (i.e., a non-additive measure or fuzzy measure), it was later on introduced by Choquet [24]. Yager proposed a generalized version in [25].

Definition 1 (Fuzzy measure) *Let $X = \{x_1, x_2, \ldots, x_n\}$ be a finite set. A discrete fuzzy measure (also called capacity) is a set function $\mu : 2^X \to [0,1]$ which is monotonic ($\mu(A) \leq \mu(B)$ for $A \subseteq B \subseteq X$) and normalized ($\mu(\emptyset) = 0$ and $\mu(X) = 1$). A fuzzy measure μ is called additive if $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \subseteq X$ such that $A \cap B = \emptyset$. Obviously, in the case of an additive measure, $\mu(A)$ is simply obtained as follows:*

$$\mu(A) = \sum_{i \in A} \mu(\{i\}) \tag{1}$$

Definition 2 (Choquet integral) *Let μ be a fuzzy measure on $X = \{x_1, x_2, \ldots, x_n\}$. The discrete Choquet integral of a function $f : X \to \mathbb{R}_+$ with respect to μ is defined as follows:*

$$C_\mu(f) = \sum_{i=1}^{n} \left(f(x_{(i)}) - f(x_{(i-1)}) \right) \cdot \mu(A_{(i)}) \ ,$$

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

Definition 3 (Möbious transform) *The Möbius transform $\mathbf{m}_\mu$ of a fuzzy measure μ is defined as follows:*

$$\mathbf{m}_\mu(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \mu(B)$$

for all $A \subseteq X$.

As a useful property of the Möbius transform, that we shall exploit later on for learning Choquet integrals, we mention that is allows for reconstructing the underlying fuzzy measure:

$$\mu(B) = \sum_{A \subseteq B} \mathbf{m}(A)$$

for all $B \subseteq X$. More specifically, we shall make use of the following representation of the Choquet integral:

$$\begin{aligned}
C_\mu(f) &= \sum_{i=1}^{n} \left(f(x_{(i)}) - f(x_{(i-1)}) \right) \cdot \mu(A_{(i)}) \\
&= \sum_{i=1}^{n} f(x_{(i)}) (\mu(A_{(i)}) - \mu(A_{(i+1)})) \\
&= \sum_{i=1}^{n} f(x_{(i)}) \sum_{R \subseteq T_{(i)}} \mathbf{m}(R) \\
&= \sum_{T \subseteq X} \mathbf{m}(T) \times \min_{(i) \in T} f(x_{(i)})
\end{aligned} \tag{2}$$

where $T_{(i)} = \{S \cup \{(i)\} \mid S \subseteq \{(i+1), \ldots, (n)\}\}$.

Definition 4 (k-Additivity) *A fuzzy measure μ is said to be k-order additive or simply k-additive if k is the smallest integer such that $\mathbf{m}(A) = 0$ for all $A \subseteq X$ with $|A| > k$.*

Thus, while a Choquet integral is determined by 2^n coefficients in general, the k-additivity of the underlying measure reduces the number of required coefficients to at most

$$\sum_{i=1}^{k} \binom{n}{i} \quad .$$

The (discrete) Choquet integral is often used as an aggregation operator, namely to aggregate the assessments $f(x_i)$ of an object on different criteria x_i into a single evaluation. If the underlying measure μ is additive (i.e., k-additive with $k = 1$), the Choquet integral reduces to a linear aggregation

$$C_\mu(f) = \sum_{i=1}^{n} w_i \cdot f(x_i) \quad ,$$

with $w_i = \mu(\{x_i\})$ the weight or, say, the importance of the criterion x_i. Besides, in this case, there is obviously no interaction between the criteria x_i, i.e., the influence of evaluation $f(x_i)$ on the overall assessment is independent of the other values $f(x_j)$, $j \neq i$.

Measuring the importance of a criterion x_i becomes obviously more involved if 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 *Shaply value* (or importance index) of x_i is defined as follows:

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

The Shaply value of μ is the vector $\varphi(\mu) = (\varphi(1), \ldots, \varphi(n))$. 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 [26], is defined as follows:

$$I(x_i, x_j) = \sum_{A \subseteq X \setminus \{x_i, x_j\}} \frac{\mu(A \cup (\{x_i, x_j\}) - \mu(A \cup (\{x_i\})) - \mu(A \cup (\{x_j\}) + \mu(A)}{(n-1)\binom{n-2}{|A|}}$$

This index ranges between -1 and 1 and indicates a positive (negative) interaction between criteria x_i and x_j if $I(x_i, x_j) > 0$ ($I(x_i, x_j) < 0$).

Interestingly, the Shaply value can also be expressed in terms of the interaction index:

$$\varphi(x_i) = \mathbf{m}(\{x_i\}) + \frac{1}{2} \sum_{x_j \in X \setminus \{x_i\}} I(x_i, x_j)$$

4 Multipartite and Object Ranking

As mentioned earlier, different types of ranking problems have recently been studied in the machine learning literature. Here, we are specifically interested in so-called *object ranking* and *multipartitle ranking*. In both problems, the goal is to learn a *ranking function* that accepts a subset $\mathcal{O} \subset \mathbf{O}$ of objects as input, and produces as output a ranking (total order) $\succeq$ of these objects. Typically, a ranking function of that kind is implemented by means of a scoring function $U : \mathbf{O} \rightarrow \mathbb{R}$, so that

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

for all $\boldsymbol{o}, \boldsymbol{o}' \in \mathbf{O}$. Obviously, $U(\boldsymbol{o})$ can be considered as a kind of utility degree assigned to the object $\boldsymbol{o} \in \mathbf{O}$. Seen from this point of view, the goal in object and multipartite ranking is to learn a latent utility function on a reference set $\mathbf{O}$. 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., that ties $U(\boldsymbol{o}) = U(\boldsymbol{o}')$ do either not occur or are broken at random.

The difference between the two problems is the type of training data available for learning such a function, and the way in which a prediction is evaluated. In object ranking, the ground truth is supposed to be a total order $\succ^*$ on $\mathbf{O}$, and training data consists of pairwise preferences of the form $\boldsymbol{o}_i \succ \boldsymbol{o}_j$. Given a new set of objects $\mathcal{O}$ to be ranked, the predicted order $\succ$ is then compared with the true order $\succ^*$ (restricted to $\mathcal{O}$). This can be done, for example, by means of a rank correlation measure such as Kendall's tau [27].

In multipartite ranking, the ground truth is supposed to be an ordinal categorization of the objects. That is, each object $\boldsymbol{o} \in \mathbf{O}$ belongs to one of the classes in $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$. Correspondingly, training data consists of labeled objects $(\boldsymbol{o}_i, \ell_i) \in \mathbf{O} \times \mathcal{L}$. Assuming that the classes are sorted such that $\lambda_1 < \lambda_2 < \ldots < \lambda_k$, the goal is to learn a ranking function $U(\cdot)$ that agrees well with this sorting 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 suitable performance measure:

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

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

$$S(u, v) = \begin{cases} 1 & u < v \\ 0 & u > v \end{cases} \tag{3}$$

indicates whether or not a pair of objects has been ranked correctly.

5 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 $\boldsymbol{o} \in \mathbf{O}$ are represented as feature vectors

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

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

$$U(\boldsymbol{o}) = C_\mu(f_{\boldsymbol{o}}) \ . \tag{4}$$

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

- The representation (4) covers the commonly used linear utility functions as a special case.
- Generalizing beyond the linear case, however, it is also able to capture more complex, non-linear dependencies and interactions between criteria.
- The Choquet integral offers various means for explaining and understanding a utility function, including the importance value and the interaction index.
- As opposed to many other models used in machine learning, the Choquet integral guarantees monotonicity in all criteria. 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 $\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_N\} \subset \mathbf{O}$, together with their feature representations $f_{\boldsymbol{o}_i}$ $(i = 1, \ldots, N)$ and a subset D of pairwise preferences between these objects; each pairwise preference is represented by a tuple $(\boldsymbol{o}_i, \boldsymbol{o}_j) \in D$, suggesting that $\boldsymbol{o}_i \succ \boldsymbol{o}_j$. While these preferences are given directly in the case of object ranking, they can be derived from the class information in the case of multipartite ranking: $(\boldsymbol{o}_i, \boldsymbol{o}_j) \in D$ if the original training data contains $(\boldsymbol{o}_i, \ell_i)$ and $(\boldsymbol{o}_j, \ell_j)$, and $\ell_i > \ell_j$.

Following the idea of empirical risk minimization [9], we seek to induce a Choquet integral that minimizes the number of ranking errors (3) 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 (2) 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:

$$\max_{M, \xi_1, \ldots, \xi_N} \left\{ M - \frac{\gamma}{|D|} \sum_{(\boldsymbol{o}_s, \boldsymbol{o}_t) \in D} \xi^s + \xi^t \right\}$$

s.t.

$$\begin{aligned}
& C_\mu(f_{\boldsymbol{o}_s}) - C_\mu(f_{\boldsymbol{o}_t}) > M - \xi^s - \xi^t && \forall (\boldsymbol{o}_s, \boldsymbol{o}_t) \in D \\
& \xi^s \geq 0 && \forall s \in \{1, \ldots, N\} \\
& \sum_{T \subseteq X} \mathbf{m}(T) = 1 && \\
& \sum_{B \subseteq A} \mathbf{m}(B) \geq 0 && \forall A \subseteq X \\
& \sum_{L \subseteq A} \mathbf{m}(L) \leq \sum_{K \subseteq B} \mathbf{m}(K) && \forall A \subset B \subseteq X
\end{aligned}$$

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 non-negative, and a positive slack is penalized in proportion to its size. Finally, γ is a trade-off parameter that controls the flexibility of the model; the higher γ, the stronger the slacks are punished.

The last three constraints formalize, respectively, the normalization, non-negativity and monotonicity of the Möbius transform. Obviously, the non-negativity 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:

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

This representation reduces the number of constraints to $n2^n$, 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 $\mathbf{m}(A) = 0$ for all $A \subseteq X$ with $|A| > k$. In fact, choosing a $k \ll n$ 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.

6 Experimental Results

We conucted experiments using a data set that classifies 172 scientific journals in the field of pure mathematics into categories A^*, A, B and C [21]. Each journal is moreover scored in terms of 5 criteria, namely

- cites: the total number of citations per year;
- IF: the well-known impact factor (average number of citations per article within two years after publication);
- II: the immediacy index measures how topical the articles published in a journal are (cites to articles in current calendar year divided by the number of articles published in that year);
- articles: the total number of articles published;
- half-line: cited half-life (median age of articles cited).

6.1 Comparison with Linear and Polynomial Kernel Methods

In a first study, we compared our approach with kernel-based methods for ranking, using the RankSVM approach with a linear and a polynomial kernel [28]. 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 resemblence with a Choquet integral. In fact, using a polynomial kernel of degree k on the original feature representation of objects, i.e., a kernel of the form

$$K(\boldsymbol{o}, \boldsymbol{o}') = (\langle f_{\boldsymbol{o}}, f_{\boldsymbol{o}'} \rangle + \lambda)^k \quad , \tag{5}$$

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 $\leq k$. Thus, a polynomial kernel of degree k captures the same interactions between criteria as a Choquet integral on a k-additive fuzzy measure.

We use an experimental setup that randomly splits the data into two parts, one half for training and one half for testing. From the training data, a total number of T object pairs is sampled, and the corresponding preferences are used for training. The model induced from this training data is then evaluated on the test data, using the C-index as a performance measure. This procedure is repeated 100 times, and the results are averaged.

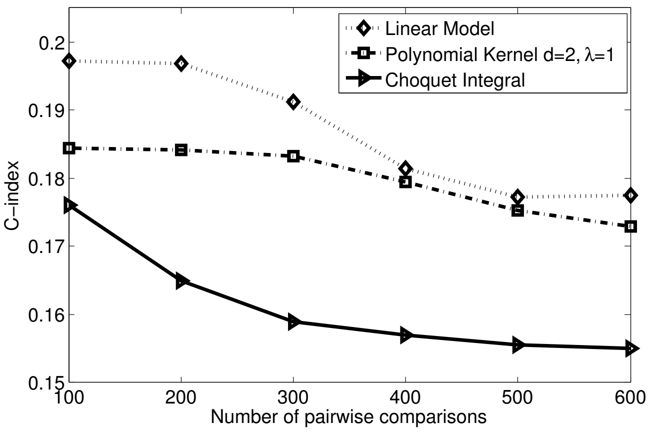

Figure 1: Average test accuracy for Choquet intregral, the linear model and the polynomial kernel of order 2.

Fig. 1 shows the average accuracy of the Choquet integral[1], the linear model and the polynomial kernel (5) with parameters $k = 2$ and $\lambda = 1$ as a function of the size T of the training set. As can be seen, the linear model performs the worst, suggesting the presence of important interactions between criteria. The kernel method is slightly better, but the best results are obtained by the Choquet integral.

6.2 Choquet Integral on k-Additive Fuzzy Measures

In a second experiment, we applied the Choquet integral with k-additive fuzzy measures, varying the value of k from 1 to 5. As can be seen in Fig. 2, there is a significant increase

[1] The trade-off parameter γ was set to 1.

cites	IF	II	articles	half-life
0.0989	0.1643	0.5379	0.0984	0.1006

Table 1: Importance of criteria in terms of the Shaply value.

in performance when going from $k = 1$ (the linear model) to $k = 2$. Increasing k beyond the value 2, however, does not seem to be beneficial.

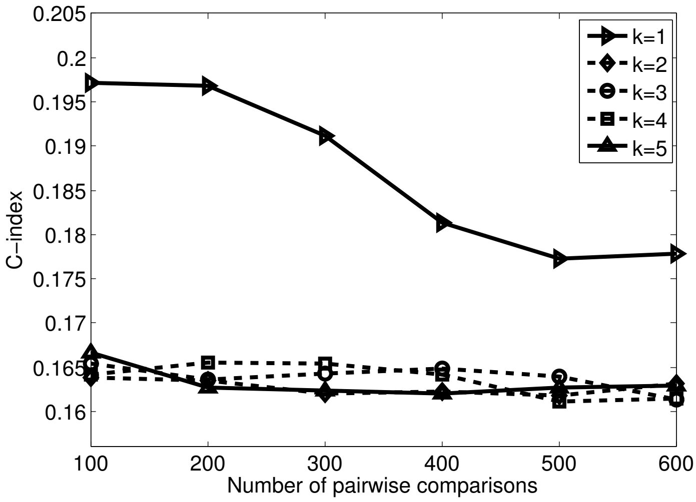

Figure 2: Performance of the Choquet integral on a k-additive fuzzy measure.

6.3 Importance of Criteria and Interaction

As mentioned before, the Choquet integral does also offer interesting information about the importance of individual criteria and the interaction between them. In fact, in many practical applications, this type of information is at least as important as the predictive accuracy of the model.

Table 1 shows the importance of the five criteria in terms of the Shaply value. As can be seen, the immediacy index and the impact factor seem to have the strongest impact on the assessment of a journal, which is hardly surprising. However, the weight of the former is even much higher than the weight of the latter, which is arguably less expected.

Table 2 shows the measures of pairwise interaction between the criteria. Interestingly, the interaction is positive throughout, i.e., there seems to be a kind of synergy between each pair of criteria. Moreover, while the degree of interaction is in general quite comparable accross all pairs of criteria, it is again maximal for the impact factor and immediacy index.

7 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 two types of ranking problems, namely object ranking and

	IF	II	articles	half-line
cites	0.24	0.26	0.34	0.29
IF		0.40	0.32	0.24
II			0.26	0.26
articles				0.29

Table 2: Pairwise interaction between criteria.

multipartite ranking. This idea is motivated by several appealing properties offered by the Choquet integral, including its ability to capture dependencies between criteria and to obey natural monotonicity conditions, as well as its interpretability.

Algorithmically, our approach is inspired by large margin methods that have been developed in the field of kernel-based machine learning. First experimental studies, in which we applied this approach to a journal ranking data set and compared it to a kernel-based ranking method, are quite promising.

Needless to say, this study is only a first step and should be complemented by more extensive experiments including diverse types of data sets. Another problem to be addressed in future work concerns the (soft) margin maximization problem. In fact, due to the large number of constraints that have to be satisfied, this problem may become computationally complex. Dedicated techniques for solving it in a more efficient way are therefore desirable.

References

[1] Fürnkranz, J.; Hüllermeier, E. (Hg.): *Preference Learning*. Springer-Verlag. 2010.

[2] Cohen, W.; Schapire, R.; Singer, Y.: Learning to Order Things. *Journal of Artificial Intelligence Research* 10 (1999), S. 243–270.

[3] Har-Peled, S.; Roth, D.; Zimak, D.: Constraint classification for multiclass classification and ranking. In: *Advances in Neural Information Processing Systems 15 (NIPS-02)* (Becker, S.; Thrun, S.; Obermayer, K., Hg.), S. 785–792. 2003.

[4] Fürnkranz, J.; Hüllermeier, E.: Pairwise Preference Learning and Ranking. In: *Proc.* ECML–03*, 13th European Conference on Machine Learning*. Cavtat-Dubrovnik, Croatia. 2003.

[5] Dekel, O.; Manning, C.; Singer, Y.: Log-linear models for label ranking. In: *Advances in Neural Information Processing Systems*. 2003.

[6] Fürnkranz, J.; Hüllermeier, E.; Vanderlooy, S.: Binary Decomposition Methods for Multipartite Ranking. In: *Proceedings ECML/PKDD–2009, European Conference on Machine Learning and Knowledge Discovery in Databases*. Bled, Slovenia. 2009.

[7] Bakir, G.; Hofmann, T.; Schölkopf, B.; Smola, A.; Taskar, B.; Vishwanathan, S. (Hg.): *Predicting structured data*. MIT Press. 2007.

[8] Choquet, G.: Theory of Capacities. *Annales de l'Institut Fourier* 5 (1954), S. 131–295.

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

[10] Grabisch, M.; Roubens, M.: *Fuzzy measures and integrals*. Physica. 2000.

[11] Grabisch, M.: Fuzzy integral in multicriteria decision making. *Fuzzy set and system* 69 (1995), S. 279–298.

[12] Torra, V.: Learning aggregation operators for preference modeling. In: *Preference Learning* (Fürnkranz, J.; Hüllermeier, E., Hg.). 2010.

[13] Torra, V.; Narukawa, Y.: *Modeling decisions: Information fusion and aggregation operators*. Spinger. 2007.

[14] Grabisch, M.: Modelling data by the Choquet integral. In: *Information Fusion in Data Mining* (Torra, V., Hg.), S. 135–148. Physica. 2003.

[15] Grabisch, M.: A new algorithm for identifying fuzzy measures and its application to pattern recognition. In: *Int. Joint Conf. of the 4th IEEE Int. Conf. on Fuzzy Systems and the 2nd Int. Fuzzy Engineering Symposium*, S. 145–150. 1995.

[16] Murofushi, T.; Mori, T.: An Analysis of Evaluation Model Using Fuzzy Measure and the Choquet Integral. In: *The 5th Fuzzy System Symposium*, S. 207–212. Kobe, Japan. 1989.

[17] Ishii, Y.; Murofushi, T.: Identification of fuzzy measures using real valued GA and considering outliers. In: *The 6th Workshop on Evaluation of Heart and Mind*. Japan. 2001.

[18] Imai, H.; Miyamori, M.; Miyakosi, M.; Sato, Y.: An algorithm based on alternative projections for a fuzzy measures identification problem. In: *International Conference on Soft Computing*. Iizuka, Japan. 2000.

[19] Imai, H.; Asano, D.; Sato, Y.: An algorithm based on alternative projections for a fuzzy measures identification problem. In: *Information Fusion in Data Mining* (Torra, V., Hg.), S. 149–159. Springer. 2003.

[20] Angilella, S.; Greco, S.; Matarazzo, B.: Non-additive robust ordinal regression with Choquet integral, bipolar and level dependent Choquet integrals. In: *IFSA-EUSFLAT*. Lisbon, Portugal. 2009.

[21] Beliakov, G.; James, S.: Citation-based journal ranks: the use of fuzzy measures. *Fuzzy Sets and Systems* (2010).

[22] Angilella, S.; Greco, S.; Matarazzo, B.: The most representive fuzzy measure for non-additive robust ordinal regression. In: *IPMU*. Dortmund, Germany. 2010.

[23] Vitali, G.: Sulla definizione di integrale delle funzioni di una variabile. *Annali di Matematica Serie IV Tomo* 2 (1925), S. 111–121.

[24] Choquet, G.: Theory of capacities. *Annales de l'Institut Fourier* 5 (1954), S. 131–295.

[25] Yager, R. R.: Generalized OWA aggregation operators. *Fuzzy Optimization and Decision Making* 3 (2004), S. 93–107.

[26] Murofushi, T.; Soneda, S.: Techniques for reading fuzzy measures (III): Interaction index. In: *The 9th Fuzzy Systems Symposium*, S. 693–696. Sapporo, Japan. 1993.

[27] Kendall, M.: *Rank correlation methods*. London: Charles Griffin. 1955.

[28] Herbrich, R.; Graepel, T.; Obermayer, K.: Large Margin Rank Boundaries for Ordinal Regression. In: *Advances in Large Margin Classifiers*, S. 115–132. 2000.