Integrating expert knowledge into kernel-based preference models

Willem Waegeman[1], Bernard De Baets[2], and Luc Boullart[1]

[1] Department of Electrical Energy, Systems and Automation, Ghent University, Technologiepark 913, B-9052 Ghent, Belgium
Willem.Waegeman@UGent.be
[2] Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

Abstract. In multi-criteria decision making (MCDM) and fuzzy modeling, preference models are typically constructed by interacting with the human decision maker (DM). When the DM experiences difficulties to specify precise for all parameters of the model, inference and elicitation procedures can assist him/her to find a satisfactory model and to assess unlabelled examples. In a related but more statistical way, machine learning algorithms can also infer preference models with similar setups and purposes, but here less interaction with the DM is integrated. We present a hybrid approach that combines the best of both worlds. It consists of a general kernel-based framework for constructing and inferring preference models, in which expert knowledge can be included. Additive models, for which interpretability is preserved, and utility models can be considered as special cases. Besides generality, important benefits of this approach are its robustness to noise and good scalability. We show in detail how this framework can be utilized to aggregate single-criterion outranking relations, resulting in a flexible class of preference models for which domain knowledge can be specified by a DM.

1 Introduction

In many situations humans compare items or objects in order to select an appropriate one for a specific goal. Think for example of buying clothes, listening to music, the dish one orders in a restaurant, etc. Continually, we evaluate objects on criteria such as appropriateness, beauty, correctness, etc. As a consequence of the growing amount of human preference information that comes available due to the fast rise of information retrieval, e-commerce and other internet-related applications, the demand for intelligent systems capable of representing and processing this information also increases. In research areas like decision making, preference modelling, fuzzy modelling, statistics and machine learning, scientists have proposed various ways to characterize such systems. In decision making and (fuzzy) preference modelling, preferences are typically modelled in a logical way, which gives the DM insights into the model and allows for the incorporation of domain knowledge into the model by interacting with the DM and the data analyst. One distinguishes input-oriented preference information (such as the relative importance of features or known values for certain parameters of the model) and result-oriented preference information (which typically consists of preference assignments by the DM for a small subset of the data, i.e. labelled training data). The machine learning community on the other hand has mainly focussed on processing these labelled training samples, and learning-based preference models are rather constructed in a statistical than a logical way, in which the uncertainty of preferences is expressed in terms of probabilities and errors instead of membership degrees or degrees of relationship. Settings like ordinal regression, ranking learning and preference learning can also capture human preference behavior, but in general less domain knowledge in terms of input-oriented information is integrated. These methods can be seen as a tool to extract and represent preference information automatically [Boutilier et al., 2004, Doyle, 2004, Jung et al., 2005]. In recent years, the decision making community has realized the need for automatic systems to support the decision making process. Inference and elicitation procedures are becoming more and more popular in this community as well, but the setup is usually slightly different. During elicitation, the DM interacts with the system and he/she can change his/her opinion about certain data objects to which labels were assigned. Moreover, he/she can partially understand the models and is able to specify values for some of the parameters. On the other hand, most elicitation procedures can only handle consistent data, i.e. noise-free data is required, and the methods usually scale badly in the number of training instances.

With this article we want to combine the best of both worlds and illustrate that these two fields have much more in common than one would expect at first sight. In particular, we will promote MCDM methods as a tool to elicit preference information from the DM that can be integrated into kernel-based machine learning algorithms. In this way, we are able to include expert knowledge into the algorithms as a kind of preprocessing step. From an MCDM point of view, this corresponds to a new way of aggregating preference relations expressed on individual features. The article is organized as follows. In Section 2 we briefly discuss the various ways of modelling preferences in machine learning and MCDM, and we pay specific attention to the aspect of inferring the parameters of the resulting models. Subsequently, in Section 3 we first give an introduction to kernel-based preference learning, followed by our framework which can be seen as a generalization of the existing approach. We show how single-criterion outranking relations can be aggregated with kernels and be embedded into this framework. At the end, we give a small example to illustrate some ideas and we formulate a conclusion.

2 Preference elicitation versus preference learning

Let us consider a (possibly infinite) set of data objects $\mathcal{X}$. We will assume that each data object $\mathbf{x} \in \mathcal{X}$ is represented by its scores on m features, which will be denoted $\mathbf{x} = (x^{(1)}, ..., x^{(m)})$, so $\mathcal{X} \equiv \mathbb{R}^m$ We remark that in MCDM data objects and features are respectively called alternatives and criteria, but here we will mainly employ a machine learning terminology. Using these notations, one can distinguish two main types of models in decision making for modelling preferences [Özturk et al., 2003]. On the one hand, we have ranking or utility models, which typically construct a continuous function of the form $f : \mathcal{X} \rightarrow \mathbb{R}$ such that:

$$\mathbf{x}_1 \succeq \mathbf{x}_2 \Leftrightarrow f(\mathbf{x}_1) \geq f(\mathbf{x}_2) \,.$$

This means that data object $\mathbf{x}_1$ is preferred to data object $\mathbf{x}_2$ if the highest value was assigned to $\mathbf{x}_1$. The ranking or utility approach has been especially popular in machine learning for scalability reasons. On the other hand, we have pairwise preference models in which the preferences are modelled by one (or more) relations $R : \mathcal{X}^2 \rightarrow [0,1]$ that express whether $\mathbf{x}_1$ should be preferred over $\mathbf{x}_2$. One can distinguish different kinds of relations such as crisp relations, fuzzy relations or reciprocal relations. Pairwise preference models allow a flexible and interpretable description of preferences and have therefore been popular in MCDM and the fuzzy set community. In this article we will mainly focus on these models, since ranking models can be reduced to pairwise preference models (under certain conditions). A pairwise preference model can for example be generated with reciprocal relations, i.e. relations $Q : \mathcal{X}^2 \rightarrow [0,1]$ satisfying

$$Q(\mathbf{x}_1, \mathbf{x}_2) + Q(\mathbf{x}_2, \mathbf{x}_1) = 1 \,, \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \,.$$

The semantics underlying reciprocal preference relations is often probabilistic: $Q(\mathbf{x}_1, \mathbf{x}_2)$ expresses the probability that object $\mathbf{x}_1$ is preferred to $\mathbf{x}_2$. One can in general construct such a reciprocal or probabilistic preference relation from a ranking model in the following way:

$$Q(\mathbf{x}_1, \mathbf{x}_2) = g(f(\mathbf{x}_1), f(\mathbf{x}_2)) \,, \tag{1}$$

with $g : \mathbb{R}^2 \rightarrow [0,1]$ usually increasing in its first argument and decreasing in its second argument [Switalski, 2003]. Examples of reciprocal preference models are Bradley-Terry models [Bradley and Terry, 1952, Agresti, 2002] and Thurstone-Case5 models [Thurstone, 1927]. They have been applied in a machine learning learning context in [Chu and Ghahramani, 2005, Herbrich et al., 2007, Radlinski and Joachims, 2007]

Another interesting type of $[0,1]$-valued preference relations are outranking relations [Perny, 1992]. They can for example be found in ELECTRE and PROMETHEE methods [Roy, 1991, Vincke, 1992, Bouyssou et al., 2006] and they have a different semantics than reciprocal relations: outranking relations are fuzzy preference relations, in the sense that they can be interpreted as graded versions of the relation $\succeq$. Global outranking relations $S : \mathbb{R}^m \times \mathbb{R}^m \rightarrow [0,1]$ are typically constructed by taking a weighted sum of single-criterion outranking relations $s_k : \mathbb{R} \times \mathbb{R} \rightarrow [0,1]$:

$$S(\mathbf{x}_1, \mathbf{x}_2) = \sum_{k=1}^{m} w_k s_k(x_1^{(k)}, x_2^{(k)}) \,. \tag{2}$$

See for example [Gheorghe et al., 2004, 2005] for a detailed discussion on that topic. Analogously, one can fuzzify the strict preference relation P, indifference relation I and incomparability relation J, leading to the concept of a (fuzzy) preference structure [De Baets et al., 1995]. Such a triplet of $[0, 1]$-valued relations can be generated from an outranking relation and an indifference generator [Van De Walle et al., 1998, De Baets and Fodor, 2003]. When no incomparability is assumed, a close relationship exists between outranking relations and reciprocal relations, although they express different concepts [De Baets and De Meyer, 2005].

In decision making one often assumes that the parameters of preference models are found by interacting with the DM, which can be a hard task. Many models contain a lot of parameters, especially when the number of criteria is rather large, and in such cases the DM often has difficulties in specifying precise values for many of the parameters. Then, inference procedures are used to infer the unknown parameters of the model from holistic judgements given by the DM. In these approaches, the information obtained from the DM can be divided into input-oriented and result-oriented information. The input-oriented information includes all domain knowledge to construct the model, such as the relative importance of criteria or the values of some of the parameters of the model. On the other hand, the result-oriented information consists of preference judgements expressed by the DM for a small subset of the alternatives. The remaining parameters are then inferred with the help of an optimization algorithm [Mousseau, 2005, Greco et al., 2008].

We will use the symbol $\mathcal{Y}$ to denote the set of judgements the DM can give and D to denote the collection of result-oriented information, which can be subdivided into ordinal class assignments and pairwise preference judgments. We will only consider the latter type of result-oriented information because under certain restrictions ordinal class assignments can be transformed to pairwise preference judgments. In this case, the DM gives crisp preference judgements about pairs of alternatives, leading to three possible choices, i.e. $\mathcal{Y} = \{+1, 0, -1\}$, in which $+1$ denotes that the first alternative is preferred to the second one, -1 denotes the opposite and 0 denotes indifference:

$$\begin{aligned} y_{ij} &= +1 \quad \Leftrightarrow \quad \mathbf{x}_i \succ \mathbf{x}_j \,, \\ y_{ij} &= \ \ 0 \quad \Leftrightarrow \quad \mathbf{x}_i \sim \mathbf{x}_j \,, \\ y_{ij} &= -1 \quad \Leftrightarrow \quad \mathbf{x}_i \prec \mathbf{x}_j \,. \end{aligned}$$

We will denote y_{ij} the label of the couple $(\mathbf{x}_i, \mathbf{x}_j)$. So, the data set $D \subseteq \mathcal{X}^2 \times \mathcal{Y}$ contains couples of alternatives together with their labels: $((\mathbf{x}_i, \mathbf{x}_j), y_{ij})$.

Inference from holistic judgements has for example been investigated by Mousseau and Slowinski [1998], Mousseau et al. [2001], Dias and Mousseau [2006] in the context of ELECTRE methods, by Fan et al. [2002], Wang and Parkan [2005] for reciprocal preference relations and by Kojadinovic [2004], Marichal et al. [2005], Grabisch et al. [2008] for Choquet integrals. These authors in general use mathematical programming techniques like linear or quadratic programs to infer the parameters of the respective models. The holistic judgements specified by the DM impose constraints on the parameters of the model, while the objective function to be minimized represents an error function. In machine learning, similar techniques are used to achieve the same goals, but the general assumptions and the mathematical representation of preferences differ. This subfield covers the ordinal regression, ranking and preference learning settings, which substantially differ from standard classification and (metric) regression learning. Ranking and ordinal regression both refer to learning utility functions, in which the result-oriented information takes the form of ordinal class assignments. In the former case, one is only interested in imposing a complete order on $\mathcal{X}$ given the result-oriented information, similar to ranking in MCDM. In the latter case, all elements of $\mathcal{X}$ have to be classified into one of the r ordinal classes. This corresponds to sorting in MCDM. Ordinal regression can be realized from the utility function, either directly from inference or by applying post-processing techniques such as ROC analysis [Waegeman et al., 2008a,b]. Ordinal regression differs from multi-class (nominal) classification because there is a linear order relation defined on $\mathcal{Y}$. Thirdly, preference learning usually refers to the situation where the DM has given pairwise preference judgements instead of ordinal class assignments [Fürnkranz and Hüllermeier, 2003] and in machine learning it has been considerably less studied than ranking and ordinal regression. Here the parameters of a reciprocal preference relation Q are inferred. A close relationship with ranking and ordinal regression models is also retained due to the transitivity of Q.

While MCDM has primarily concentrated on preference modelling for traditional decision making problems, applications of preference learning are rather found in more fancy domains with different characteristics (less domain knowledge, less interaction, more noise, higher-dimensional data, etc.). Nevertheless, the idea of inferring the parameters of a preference model from holistic judgements remains a common goal and a cross-fertilization between both fields could be meaningful. In the next section we present a method that has a flavor of both fields, namely the robustness to noise, scalability and generality of machine learning methods and the adequate way of including domain knowledge from MCDM.

3 Kernel methods for preference modelling

During the last decade, a lot of interesting papers on ordinal regression, ranking and preference learning have appeared, see e.g. [Herbrich et al., 2000, Freund et al., 2003, Crammer and Singer, 2001, Shashua and Levin, 2003, Rennie and Srebro, 2005, Chu and Keerthi, 2005]. Many of these authors use kernel methods to design learning algorithms. The majority of them also considers ranking approaches to represent the preferences. We first briefly explain the ordinal regression approach introduced by Herbrich et al. [2000] to model preferences, followed by our own approach based on pairwise preference models, which can be seen as a generalization of the ranking approach. In the last subsections we demonstrate how this method could be useful to aggregate single-criterion preference relations and we give an example.

3.1 Ranking models

In the context of kernel-based ordinal regression or ranking, we consider ranking functions $f : \mathcal{X} \to \mathbb{R}$ of the following general form:

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x})\,, \tag{3}$$

with ϕ a possibly infinite-dimensional and in general unknown feature mapping. Given a dataset

$$D = \{((\mathbf{x}_i, \mathbf{x}_j), y_{ij}) \mid i, j \in \mathbb{N}^*, \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}\}\,,$$

consisting of pairwise preference assignments given by a DM, the parameters $\mathbf{w}$ are inferred by solving a quadratic program defined as follows:

$$\min_{\mathbf{w},\xi_{ij}} \quad \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i,j:y_{ij}=1} \xi_{ij}$$

$$\text{subject to} \begin{cases} \forall i, j : y_{ij} = 1 : \mathbf{w} \cdot (\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)) \geq 1 - \xi_{ij} \\ \xi_{ij} \geq 0\,. \end{cases}$$

Similar to other kernel-based learning algorithms like support vector machines [Cristianini and Shawe-Taylor, 2000], the function to be optimized puts together two objectives in a weighted sum with C a trade-off parameter. The first objective is the standard regularization criterion used in kernel methods. Such a regularizer can be interpreted as a functional representing the complexity of the ranking function in Eq. (3). In the dual formulation this regularizer is rewritten in terms of kernels and must be interpreted as the norm of f in the feature space $\mathcal{H}$. The second objective reflects an error function defined on the difference between the holistic judgements and the output of the model. To this end, a slack variable ξ_{ij} is introduced for each pair of better/worse objects. We are thus looking for a utility model of type Eq. (3) that preserves the preferences given by the DM as good as possible, while simultaneously the complexity of the model is bounded.

In the dual formulation, the feature mappings in the optimization problem can be expressed in terms of kernels:

$$\max_{\alpha_{ij}} \sum_{i,j:y_{ij}=1} \alpha_{ij} - \frac{1}{2} \sum_{i,j:y_{ij}=1} \sum_{k,l:y_{kl}=1} \alpha_{ij}\alpha_{kl}(K(\mathbf{x}_i, \mathbf{x}_k) - K(\mathbf{x}_i, \mathbf{x}_l) - K(\mathbf{x}_j, \mathbf{x}_k) + K(\mathbf{x}_j, \mathbf{x}_l))\,,$$

$$\text{subject to } 0 \leq \alpha_{ij} \leq C, \quad \forall i, j : y_{ij} = 1\,.$$

and the model becomes

$$f(\mathbf{x}) = \sum_{i,j:y_{ij}=1} \alpha_{ij}(K(\mathbf{x}, \mathbf{x}_i) - K(\mathbf{x}, \mathbf{x}_j))\,. \tag{4}$$

Given such a ranking model, a reciprocal preference relation can be constructed in several ways. In general it will take the form of Eq. (1).

3.2 Construction of a pure pairwise preference model

In this section we will present a generalization of the above approach. Let us now consider $\delta f : \mathcal{X}^2 \to \mathbb{R}$ expressing the preference of the object $\mathbf{x}$ compared to the object $\mathbf{x}'$:

$$\delta f(\mathbf{x}, \mathbf{x}') = \mathbf{w} \cdot \delta\phi(\mathbf{x}, \mathbf{x}') \,.$$

For such a relation, we can draw up a similar type of optimization problem, namely

$$\min_{\mathbf{w}, \xi_{ij}} \quad \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i,j:y_{ij}=1} \xi_{ij} \tag{5}$$

$$\text{subject to} \begin{cases} \forall i,j : y_{ij} = 1 : \quad \mathbf{w} \cdot \delta\phi(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \xi_{ij} \\ \xi_{ij} \geq 0 \,. \end{cases}$$

With duality theory, the entire setup is reformulated in terms of kernels. Following from the Karush-Kuhn-Tucker conditions, we get a similar dual optimization problem:

$$\max_{\alpha_{ij}} \sum_{i,j:y_{ij}=1} \alpha_{ij} - \frac{1}{2} \sum_{i,j:y_{ij}=1} \sum_{k,l:y_{kl}=1} \alpha_{ij}\alpha_{kl} K^*(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)$$

$$\text{subject to } 0 \leq \alpha_{ij} \leq C, \quad \forall i,j : y_{ij} = 1 \,,$$

in which the kernel now represents a dot-product on couples of alternatives in an unknown feature space $\mathcal{H}$:

$$K^*(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) = \delta\phi(\mathbf{x}_i, \mathbf{x}_j) \cdot \delta\phi(\mathbf{x}_k, \mathbf{x}_l) \,.$$

The solution of optimization problem (5) can be written as:

$$\delta f(\mathbf{x}, \mathbf{x}') = \sum_{i,j:y_{ij}=1} \alpha_{ij} K^*(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}, \mathbf{x}') \,, \tag{6}$$

and regularization now takes the following form:

$$||\mathbf{w}||^2 = \sum_{i,j:y_{ij}=1} \sum_{k,l:y_{kl}=1} \alpha_{ij}\alpha_{kl} K(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) \,. \tag{7}$$

Remark that the ranking approach can be seen as a special case, by defining the kernel function K^* as

$$K^*(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) = K(\mathbf{x}_i, \mathbf{x}_k) - K(\mathbf{x}_i, \mathbf{x}_l) - K(\mathbf{x}_j, \mathbf{x}_k) + K(\mathbf{x}_j, \mathbf{x}_l) \,,$$

of the original two-dimensional kernel function used in the utility model in Eq. (4).

3.3 Aggregating single-criterion outranking relations with kernels

The framework presented in the previous subsection is to our opinion very general. By picking a specific kernel function, the data analyst can choose the desired model complexity. With a linear kernel one obtains an additive model, with a polynomial kernel interactions between criteria are considered and with an RBF kernel one arrives at a full black-box representation of the data, useful for very complex domains. Moreover, this framework includes utility models as a special case. In this subsection we demonstrate how common MCDM models can be embedded into it. We will give the example of concordance relations. Similar techniques can be used to kernelize Choquet integrals for example, but that discussion would deserve an article on its own.

In particular, kernel methods can be seen as a tool for aggregating single-criterion preference relations. We will give the example of valued concordance relations, which are in ELECTRE methods constructed from single-criterion outranking relations. Following the notation adopted in this paper, we can represent such a single-criterion outranking relation as follows:

$$s_k(x_i^{(k)}, x_j^{(k)}) = \frac{p_k(x_i^{(k)}) - \min\{x_j^{(k)} - x_i^{(k)}, p_k(x_i^{(k)})\}}{p_k(x_i^{(k)}) - \min\{x_j^{(k)} - x_i^{(k)}, q_k(x_i^{(k)})\}}\,,$$

with p_k and q_k threshold functions. A concordance relation is built from these single-criterion outranking relations as a weighted sum:

$$S(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{m} w_k s_k(x_i^{(k)}, x_j^{(k)}) = \mathbf{w} \cdot \mathbf{s}(\mathbf{x}_i, \mathbf{x}_j)\,,$$

in which we used a shorthand (vector) notation in the last line. The main idea simply consists of embedding this vector $\mathbf{s}$ of single-criterion preferences in feature space:

$$\delta f(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{w} \cdot \delta\phi(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{w} \cdot \phi(\mathbf{s}(\mathbf{x}_i, \mathbf{x}_j))\,. \tag{8}$$

Remark that ϕ still represents the mapping of an m-dimensional real vector (in this case $\mathbf{s}$) to a feature space of higher dimension, while $\delta\phi$ transforms couples of alternatives to feature space. The construction of the vector of single-criterion outranking relations from $\mathbf{x}_i$ and $\mathbf{x}_j$ can thus be interpreted as a preprocessing step, preparatory to learning or inference. In this way domain knowledge is taken into account when the DM determines the values of the single-criterion outranking relations, apart from the inference procedure.

In the dual representation, the above model can be rewritten in the general form of Eq. (6) with the kernel function K^* now defined by:

$$K^*(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) = K(\mathbf{s}(\mathbf{x}_i, \mathbf{x}_j), \mathbf{s}(\mathbf{x}_k, \mathbf{x}_l))\,, \tag{9}$$

and $K : \mathcal{X}^2 \to \mathbb{R}$ a regular kernel function. The choice of K will now determine whether single-criterion outranking relations are aggregated in an additive or more complex way and, when interactions are considered, one can easily specify how many criteria simultaneously interact by specifying the degree of a polynomial kernel. Alternatively, the degree of such a kernel can be also inferred with resampling techniques.

With a further restriction to linear kernels and a normalization, one recovers the concordance relation as briefly summarized in the following proposition that directly follows from the definitions.

Proposition 1. *Let $K : \mathcal{X}^2 \to \mathbb{R}$ be a positive definite kernel with feature mapping $\phi : \mathbb{R}^m \to \mathbb{R}^{m^*}$, let ϕ_k be the k-th component of ϕ with $k \in \{1, ..., m^*\}$, let δf and K^* be respectively defined by Eq. (6) and (9) and let*

$$\sum_{k=1}^{m^*} \sum_{i,j:y_{ij}=1} \alpha_{ij} \phi_k(\mathbf{s}(\mathbf{x}_i, \mathbf{x}_j)) = 1 \,, \tag{10}$$

then δf is an outranking relation.

Eq. (10) is the equivalent representation of the normalization of $\mathbf{w}$. Such a normalization in terms of the L_1-norm can only be expressed in the primal representation, using $\delta\phi$ instead of K^*, contrary to the L_2-norm of $\mathbf{w}$ which is given in kernel form by substituting Eq. (9) into Eq. (7).

This limitation of the L_1-norm might impose a serious bottleneck on the scalability of our algorithm if δf has to be an outranking relation. Yet, the grounds for avoiding a normalization of the weights are much more fundamental and go back to the need for regularization. From a machine learning perspective, such a normalization (which fixes the complexity of the model) will lead to overfitting. Definitely, the lack of regularization is an undeniable shortcoming of existing inference procedures in MCDM.

Knowing that a normalization is better avoided, we do not promote the approach presented here as a tool to construct outranking relations. We rather want to illustrate how preferences relations defined on individual criteria can be aggregated with kernels, leading to a flexible class of models that can be efficiently and robustly inferred based on holistic judgements.

3.4 Example

The embedding of single-criterion preference into kernel space is straightforward and can be efficiently implemented in a standard SVM, a Ranking SVM and other kernel-based methods by constructing a set of positive and negative examples from the single-criterion preference relations. Unfortunately, we could not find any publicly available data sets that satisfied our needs to demonstrate the potential benefits of our approach. In the MCDM domain it is not common to report experimental results on statistically relevant samples, simply because typical applications give rise to very small data sets on which a statistical method would fail to produce meaningful results. To the best of our knowledge, no data sets of reasonable size exist, containing both preference evaluations for individual criteria and holistic judgments.

Alternatively, we will briefly discuss an example adopted from Gheorghe et al. [2004]. In that paper an artificial data sample of three data objects is considered. Each of the data objects is evaluated on three criteria. The actual values obtained for these criteria do not matter, since we will only need the single-criterion preference relations constructed from them. In the example strict preference relations are considered instead of concordance relations. Similarly, we will employ the notation $p_k(x_i^{(k)}, x_j^{(k)})$ for such a single-criterion preference relation. The following $[0,1]$-valued strict preference relations were generated for each of the three criteria:

$p_k(x_i^{(1)}, x_j^{(1)})$	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$
$x_1^{(1)}$	0	1	0.599
$x_2^{(1)}$	0	0	0
$x_3^{(1)}$	0	1	0

$p_k(x_i^{(2)}, x_j^{(2)})$	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$
$x_1^{(2)}$	0	1	0.49
$x_2^{(2)}$	0	0	0
$x_3^{(2)}$	0.056	0.956	0

$p_k(x_i^{(3)}, x_j^{(3)})$	$x_1^{(3)}$	$x_2^{(3)}$	$x_3^{(3)}$
$x_1^{(3)}$	0	0	0
$x_2^{(3)}$	1	0	0.956
$x_3^{(3)}$	1	0	0

This information is generated in a preprocessing step by interacting with the DM. We assume in addition that the DM has given the following pairwise judgments:

$$y_{12} = +1\,, \quad y_{13} = +1\,, \quad y_{23} = -1\,,$$

so transitivity is preserved and we obtain the ranking $\mathbf{x}_1 \succeq \mathbf{x}_3 \succeq \mathbf{x}_2$. This ranking is consistent with the first two criteria and inconsistent with the third one. However, if we allow that criteria can have a negative impact on the preference, then the holistic judgments are also consistent with the third criterion. Remark that the models discussed above allow negative impacts of criteria. In fact, it is supposed that the direction of the impact of a criterion will be inferred as well. If this direction is known beforehand, then one could add a constraint to the optimization problem.

In order to apply the method described above, we transform the available data to a standard classification data set as follows:

	$p_k(x_i^{(1)}, x_j^{(1)})$	$p_k(x_i^{(2)}, x_j^{(2)})$	$p_k(x_i^{(3)}, x_j^{(3)})$	y_{ij}
1	1	1	0	+1
2	0.599	0.49	0	+1
3	0	0	0.956	-1
4	0	0	1	-1
5	0	0.056	1	-1
6	1	0.956	0	+1

Subsequently, we train a support vector machine on this data set and obtain the following output values for $C = 1$ and a linear kernel:

$\delta f(\mathbf{x}_i, \mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	/	-1.000	-0.713
$\mathbf{x}_2$	1.020	/	0.995
$\mathbf{x}_3$	1.000	-1.001	/

From this output, a reciprocal preference relation can be constructed in a post-processing step. In this case, we applied the method of [Wu et al., 2004] that utilizes Eq. (1):

$Q(\mathbf{x}_i, \mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	/	0.840	0.770
$\mathbf{x}_2$	0.202	/	0.208
$\mathbf{x}_3$	0.207	0.837	/

We emphasize that the obtained probabilistic measure of preference is not well calibrated here, because this example is far too small to draw conclusions in a statistical way. Nevertheless, the probabilities still match very well with the preference assignments given by the DM, because they are consistent with the ranking $\mathbf{x}_1 \succeq \mathbf{x}_3 \succeq \mathbf{x}_2$.

4 Conclusion

In this article we presented a first attempt to bridge the gap between machine learning and MCDM in modelling human preference behavior. To this end, we generalized an existing kernel-based framework for preference learning such that expert knowledge can be integrated. This approach can be interpreted as a kind of preprocessing step in which preference relations defined on individual features are specified by the DM. Following on that, kernel method are employed to aggregate these single-criterion preference relations in a single preference model. In general, additive models are generated by a linear kernel, while more complex models are characterized by other types of kernels. We demonstrated that with polynomial kernels it is possible to represent interactions between criteria, in a very similar way as Choquet integrals for example. Due to a lack of usable data sets in this domain, we could not demonstrate the practical usefulness of our approach empirically. In the future we intend to set up some experiments to generate such data ourselves.

Acknowledgment

Willem Waegeman is supported by a grant of the "Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)".

Bibliography

A. Agresti. *Categorical Data Analysis, 2nd version*. John Wiley and Sons, 2002.

C. Boutilier, R. Brafman, C. Domshlak, H. Hoos, and D. Poole. Cp-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research*, 21:135–191, 2004.

D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukias, and P. Vincke. *Evaluation and Decision Models with Multiple Criteria*. Springer, 2006.

R. Bradley and M. Terry. Rank analysis of incomplete block designs. I: The method of paired comparisons. *Biometrika*, 39:324–345, 1952.

W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

W. Chu and S. Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the International Conference on Machine Learning, Bonn, Germany*, pages 321–328, 2005.

K. Crammer and Y. Singer. Pranking with ranking. In *Proceedings of the Conference on Neural Information Processing Systems, Vancouver, Canada*, pages 641–647, 2001.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

B. De Baets and H. De Meyer. Transitivity frameworks for reciprocal relations: cycle-transitivity versus FG-transitivity. *Fuzzy Sets and Systems*, 152:249–270, 2005.

B. De Baets and J. Fodor. *Principles of Fuzzy Preference Modelling and Decision Making*, chapter Additive Fuzzy Preference Structures, the Next Generation, pages 15–27. Academia Press, 2003.

B. De Baets, B. Van De Walle, and E. Kerre. Fuzzy preference structures without incomparability. *Fuzzy Sets and Systems*, 76:333–348, 1995.

L. Dias and V. Mousseau. Inferring ELECTRE's veto-related parameters from outranking examples. *European Journal of Operational Research*, 170:172–191, 2006.

J. Doyle. Prospects for preferences. *Computational Intelligence*, 20:111–136, 2004.

Z. Fan, J. Ma, and Q. Zhang. An approach to multiple-attribute decision making based on fuzzy preference information on alternatives. *Fuzzy Sets and Systems*, 131:101–106, 2002.

Y. Freund, R. Yier, R. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

J. Fürnkranz and E. Hüllermeier. Pairwise preference learning and ranking. *Lecture Notes in Computer Science*, 2837:145–156, 2003.

R. Gheorghe, A. Bufardi, and P. Xirouchakis. Construction of two-parameters fuzzy outranking relations from fuzzy evaluations. *Fuzzy Sets and Systems*, 143:391–412, 2004.

R. Gheorghe, A. Bufardi, and P. Xirouchakis. Construction of global fuzzy preference structures from two-parameter single-criterion fuzzy outranking relations. *Fuzzy Sets and Systems*, 153:303–330, 2005.

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. *European Journal of Operational Research*, 186:766–785, 2008.

S. Greco, V. Mousseau, and R. Slowinski. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research*, 191:415–435, 2008.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 115–132. The MIT Press, 2000.

R. Herbrich, T. Minka, and T. Graepel. Trueskill: a bayesian skill rating system. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. The MIT Press, Cambridge, MA, 2007.

S. Jung, J.H. Hong, and T.-S. Kim. A statistical model for user preference. *IEEE Transactions on Knowledge and Data Engineering*, 17:834–843, 2005.

I. Kojadinovic. Estimation of the weights of interacting criteria from the set of profiles by means of information-theoretic functionals. *European Journal of Operational Research*, 155:741–751, 2004.

J.L. Marichal, P. Meyer, and M. Roubens. Sorting multi-attribute alternatives: The TOMASO method. *Computers & Operations Research*, 32:861–877, 2005.

V. Mousseau. A general framework for constructive learning preference elicitation in multiple criteria decision aid. Technical report, LAMSADE, Université Paris-Dauphine, France, 2005.

V. Mousseau and R. Slowinski. Inferring an ELECTRE TRI model from assignment examples. *Journal of Global Optimization*, 12:157–174, 1998.

V. Mousseau, J. Figueira, and J. Naux. Using assignment examples to infer weights for ELECTRE TRI method: Some experimental results. *European Journal of Operational Research*, 130(2):263–275, 2001.

M. Öztürk, A. Tsoukias, and P. Vincke. Preference modelling. Technical report, DIMACS, Rutgers University, 2003.

P. Perny. *Modelisation, Agrégation et Exploitation de Préférences Floues dans une Problématique de Rangement*. PhD thesis, Université Paris-Dauphine, 1992.

F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining, San Jose, CA, U.S.A.*, pages 570–579, 2007.

J. D. Rennie and N. Srebro. Loss functions for preference levels: Regression with discrete, ordered labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling, Edinburgh, Scotland*, pages 180–186, 2005.

B. Roy. The outranking approach and the foundations of ELECTRE methods. *Theory and Decision*, 31(1):49–73, 1991.

A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. In *Proceedings of the International Conference on Neural Information Processing Systems, Vancouver, Canada*, pages 937–944. The MIT Press, 2003.

Z. Switalski. General transitivity conditions for fuzzy reciprocal preference matrices. *Fuzzy Sets and Systems*, 137:85–100, 2003.

L. Thurstone. A law of comparative judgment. *Psychological Review*, 79:281–299, 1927.

B. Van De Walle, B. De Baets, and E. Kerre. Characterizable fuzzy preference structures. *Annals of Operations Research*, 80:105–136, 1998.

P. Vincke. *Multicriteria Decision-Aid*. Wiley, 1992.

W. Waegeman, B. De Baets, and L. Boullart. ROC analysis in ordinal regression learning. *Pattern Recognition Letters*, 29:1–9, 2008a.

W. Waegeman, B. De Baets, and L. Boullart. On the scalability of ordered multi-class ROC analysis. *Computational Statistics and Data Analysis*, 52: 3371–3388, 2008b.

Y.-M. Wang and C. Parkan. Multiple attribute decision making based on fuzzy preference information on alternatives: Ranking and weighting. *Fuzzy Sets and Systems*, 153:331–346, 2005.

F. Wu, C. Lin, and R. Weng. Probability estimates for multi-class support vector machines by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.