Artificial Intelligence 175 (2011) 1223–1250

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

On the ERA ranking representability of pairwise bipartite ranking functions

Willem Waegeman *, Bernard De Baets

KERMIT, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 31 January 2009
Received in revised form 19 June 2010
Accepted 19 June 2010
Available online 2 December 2010

Keywords:
Pairwise bipartite ranking
Reciprocal preference relation
Cycle transitivity
Receiver operating characteristics (ROC) analysis
Graph theory
Multi-class classification
Decision theory
Machine learning

ABSTRACT

In domains like decision theory and social choice theory it is known for a long time that stochastic transitivity properties yield necessary and sufficient conditions for the ranking or utility representability of reciprocal preference relations. In this article we extend these results for reciprocal preference relations originating from the pairwise comparison of random vectors in a machine learning context. More specifically, the expected ranking accuracy (ERA) is such a reciprocal relation that occurs in multi-class classification problems, when ranking or utility functions are fitted to the data in a pairwise manner. We establish necessary and sufficient conditions for which these pairwise bipartite ranking functions can be simplified to a single ranking function such that the pairwise expected ranking accuracies of both models coincide. Similarly as for more common reciprocal preference relations, cycle transitivity plays a crucial role in this new setting. We first consider the finite sample case, for which expected ranking accuracy can be estimated by means of the area under the ROC curve (AUC), and subsequently, we further generalize these results to the underlying distributions. It turns out that the ranking representability of pairwisely compared random vectors can be expressed elegantly in a distribution-independent way by means of a specific type of cycle transitivity, defined by a conjunctor that is closely related to the algebraic product.

1. Introduction

Multi-class classification and ordinal regression can be seen as two closely related machine learning settings that share many properties. Multi-class classification refers to the supervised learning problem of inferring a predictive model capable of classifying data into a finite number of classes. This simply means that the model predicts for new data instances an output (also called label or response variable) that takes values in a finite unordered set (for example, class labels red, green, blue). Ordinal regression considers a slightly different setting. Labels here come from a finite ordered set, in which the order naturally follows from the semantics of the classes (for example, class labels bad, moderate, good). As a specific case of preference learning, ordinal regression problems typically arise in situations where humans are involved in the data generation process, like human experts or internet users expressing preferences on objects w.r.t. characteristics such as quality, beauty, appropriateness, etc.

So, the different semantics of the data respectively result in the absence or presence of an order relation on the classes in multi-class classification or ordinal regression. Owing to this important interpretation of the classes, substantially different methods have been proposed in the past for the two types of learning problems. Briefly summarized, the absence or presence of an order relation leads to two main differences in assumptions:

* Corresponding author. Tel.: +329 264 6018, fax: +329 264 6220.
E-mail address: Willem.Waegeman@UGent.be (W. Waegeman).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2010.11.006

(1) Firstly, both models typically differ in the type of performance measure they optimize. If an order relation on the classes can be assumed, then a performance measure that takes this order into account must be utilized, both for optimization and evaluation. For example, in ordinal regression, misclassifying an object of class "bad" into class "good" must typically lead to a higher loss than misclassifying the same object into class "moderate".
(2) Secondly, the absence or presence of an order relation on the classes gives rise to a different model structure for the two types of problems. The model structure of multi-class classification methods typically consists of an ensemble of binary classifiers, such as one-versus-one [26,30] and one-versus-all [41] ensembles, while typically only one global model is considered in ordinal regression. Moreover, this global model always consists of an underlying latent variable that reflects the order on the classes. Let $\mathcal{X}$ denote the set of data objects, then this latent variable serves as a ranking function $f : \mathcal{X} \to \mathbb{R}$ that defines a total order on the data objects. The final decision rule is then in the end obtained by placing a number of thresholds on the ranking function. This is for example the case in traditional statistical ordinal regression algorithms [2,38] and kernel-based methods [6,42].

Several authors [27,35,44] empirically analyzed in recent work the relationship between multi-class classification and ordinal regression, in which they primarily aim to improve ordinal regression algorithms by using ideas from multi-class classification, without considering an underlying ranking function. Conversely, the motivation of this article is to improve multi-class classification algorithms by using techniques from ordinal regression. Moreover, we will mainly focus on the theoretical connections between both problem settings, and to establish such a connection, we will take the ranking function that characterizes ordinal regression models as starting point. In this context, expected ranking accuracy (ERA) is a ranking-based performance measure that has recently been introduced for bipartite ranking [1] and further extended to ordinal regression [47]. Expected ranking accuracy can be easily considered too in multi-class classification, especially for one-versus-one ensembles, where the ensemble contains a set of pairwise bipartite ranking functions (i.e. one bipartite ranking function for each pair of classes). By using concepts from receiver operator characteristics (ROC) analysis, graph theory, decision theory and preference modeling, we will show that transitivity properties of the reciprocal relation generated by expected ranking accuracy result in a connection between multi-class classification and ordinal regression models.

Roughly speaking, we will investigate the conditions for which a one-versus-one ensemble, containing a set of bipartite ranking functions, can be reduced to an ordinal regression model with only one underlying ranking function, such that both models obtain an identical performance in terms of expected ranking accuracy. We will further refer to this property as ERA ranking representability of a one-versus-one ensemble. ERA ranking representability can be interpreted as a natural extension to the infinite sample case of AUC ranking representability, as previously introduced in [46]. It is well known that the area under the ROC curve (AUC) forms an unbiased estimator of the expected ranking accuracy on a finite dataset. Let us as an introductory example in a multi-class classification setting consider the following hypothetical three-class dataset that contains six objects of each class:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
y_i	$\mathcal{C}_1$	$\mathcal{C}_1$	$\mathcal{C}_1$	$\mathcal{C}_1$	$\mathcal{C}_1$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_2$	$\mathcal{C}_2$	$\mathcal{C}_2$	$\mathcal{C}_2$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_3$	$\mathcal{C}_3$	$\mathcal{C}_3$	$\mathcal{C}_3$	$\mathcal{C}_3$

We have for simplicity assigned the indices in such a way that pairwise AUCs can be computed easily for a given ranking. Remark that the AUC simply computes the fraction of (lower class, higher class) couples that are correctly ranked by the classifier. Let us suppose that the following triplet of bipartite ranking functions is statistically inferred by a one-versus-one ensemble for this small toy problem:

	i											
ranking for f_{12}	7	8	1	2	9	3	4	5	6	10	11	12
ranking for f_{23}	13	7	14	8	9	10	11	12	15	16	17	18
ranking for f_{13}	13	1	2	3	14	15	16	17	18	4	5	6

So, from left to right, the numbers represent the ranking of the indices of the data objects, respectively obtained with the ranking functions f_{12}, f_{23} and f_{13}. For the pairwise AUCs we find:

$$\widehat{A}_{12}(f_{12}, D) = 20/36, \qquad \widehat{A}_{23}(f_{23}, D) = 25/36, \qquad \widehat{A}_{13}(f_{13}, D) = 15/36. \tag{1}$$

In other words, one finds for instance that 20 of the 36 couples are correctly ranked by the ranking function f_{12}: object number 1 is ranked before four objects of class $\mathcal{C}_2$, as well as object number 2, object number 3 is ranked before three objects of class $\mathcal{C}_2$, and so on. A more formal definition of the AUC will be given in Section 2.

In this example, the triplet of bipartite rankings can still be replaced in different ways by a single ranking of the whole data set such that the same pairwise AUCs are measured, for example

	i																	
ranking for global f	13	1	2	3	7	8	9	10	11	14	15	16	17	18	4	5	12	6

is such a ranking that results in the same pairwise AUCs. Verification of AUC ranking representability is much more difficult for larger datasets, since enumerating all global rankings is then computationally infeasible. However, in [46] we have shown

that AUC ranking representability is strongly linked with a specific type of transitivity that has been called AUC transitivity for this reason. In Section 4 we will first recapitulate necessary transitivity conditions for AUC ranking representability by explaining the link between bipartite rankings and collections of dice. The reciprocal relations observed in both problems exhibit a specific type of transitivity that has been called dice transitivity [15]. Due to a specific requirement imposed for bipartite rankings, dice transitivity does not yield a sufficient condition. Because of that, we also introduced a new type of transitivity based on graph-theoretic concepts. This condition, which is called AUC transitivity, imposes constraints on the values of the pairwise AUCs, and it gives rise to a sufficient transitivity condition for AUC ranking representability. As a result, AUC transitivity can be verified by solving an integer quadratic program. Moreover, in Sections 5 and 6 a closed-form expression for the solution of this integer quadratic program will be derived, so that a combinatorial optimization procedure can be avoided.

As shown in the following sections, ERA and AUC ranking representability strongly rely on the notion of a reciprocal relation, because the ERA and the AUC can be considered as specific examples of such relations. Historically, the representability of reciprocal relations in terms of a single ranking or utility function has been extensively studied in domains like utility theory [22], preference modelling [39], social choice theory [25], fuzzy set theory [4] and mathematical psychology [18,37,45], as a characterization of human preference judgments. Especially for reciprocal preference relations $Q : \mathcal{X}^2 \to [0,1]$ on a set $\mathcal{X}$ of data objects or alternatives, it has been shown that the notion of transitivity plays a crucial role. We recall that the reciprocity property expresses that for all $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^2$ it holds that

$$Q(\mathbf{x}_1, \mathbf{x}_2) + Q(\mathbf{x}_2, \mathbf{x}_1) = 1,$$

with the assumption $Q(\mathbf{x}_1, \mathbf{x}_1) = 1/2$. The above-mentioned authors all observed that Q has to satisfy some specific transitivity conditions in order to be representable in terms of a single ranking or utility function $f : \mathcal{X} \to \mathbb{R}$ in the following sense: for any $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^2$ it holds that

$$Q(\mathbf{x}_1, \mathbf{x}_2) \leqslant \frac{1}{2} \quad \Leftrightarrow \quad f(\mathbf{x_1}) \leqslant f(\mathbf{x_2}).$$

Reciprocal preference relations for which this representation holds are called weak utility models [37]. The latter proved that a reciprocal preference relation is a weak utility model if and only if it satisfies weak stochastic transitivity, i.e., for any $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \in \mathcal{X}^3$ it holds that

$$\big(Q(\mathbf{x}_1, \mathbf{x}_2) \geqslant 1/2 \wedge Q(\mathbf{x}_2, \mathbf{x}_3) \geqslant 1/2\big) \quad \Rightarrow \quad Q(\mathbf{x}_1, \mathbf{x}_3) \geqslant 1/2. \tag{2}$$

Remark that the pairwise AUCs given in (1) do not satisfy weak stochastic transitivity. Analogous to weak utility models, one can define other, typically stronger, conditions on the relationship between Q and f, leading to stronger transitivity conditions like moderate or strong stochastic transitivity. For the utility representability of fuzzy preference relations, similar forms of transitivity exist by using t-norms [5,10,43]. As shown in the present paper, these types of transitivity and the more general umbrella of cycle transitivity [11] are valuable tools for the analysis of reciprocal preference relations as well. We will give a short and very incomplete overview of different types of transitivity in Section 3. By exploiting the graph-theoretic reformulation of the AUC and introducing a new type of transitivity, we were able to derive necessary and sufficient conditions for AUC ranking representability. Here we further extend these results by considering the infinite sample case instead of the finite sample case. As a result, we will once more introduce a new type of transitivity that can be categorized in the framework of cycle transitivity. Using this framework, we will also examine the connection with other types of transitivity that have been proposed in the context of dice games [15] and the pairwise comparison of random variables [12,14,16].

This article is organized as follows. In Section 2 the machine learning concepts mentioned in this introduction are more formally described. Subsequently, in Section 3 different existing forms of transitivity are discussed and the framework of cycle transitivity is briefly outlined. This allows us to present and extend in Section 4 necessary and sufficient conditions for AUC ranking representability by means of a new type of transitivity with the suitable name of AUC transitivity. In Section 5 we then present the most important contribution of this paper: the generalization of transitivity properties to ERA ranking representability. Finally, Section 6 discusses some practical considerations, followed by a general conclusion.

2. Expected ranking accuracy

In the last decade, the problem of ranking, i.e., statistically inferring the parameters of a ranking function $f : \mathcal{X} \to \mathbb{R}$ from a finite data set, has grown out to an active and widespread research field that covers applications like information retrieval, marketing, financial forecasting and more traditional decision making problems (see e.g. [7,9,31,34]). We will in particular focus on pairwise bipartite ranking in a multi-class setting. Such a setting basically implies that one aims to construct a statistical model that describes the relationship between data objects $\mathbf{x} \in \mathcal{X}$ on the one hand and a (usually small) unordered set of r classes $\mathcal{Y} = \{\mathcal{C}_1, \ldots, \mathcal{C}_r\}$ on the other hand. Although different methods have been proposed for extending binary classification algorithms ($r = 2$) to multi-class classification ($r > 2$), the pairwise approach [26,30] has been especially popular due to its simplicity, good performance and generality. This approach in essence fits a binary classifier to the data for each pair of classes. It is for this reason also called a one-versus-one classification scheme. Since many binary

classification methods like logistic regression, linear discriminant analysis, neural networks and support vector machines construct internally a latent continuous variable, a set $\overline{\mathcal{F}}$ of bipartite ranking functions $f_{kl} : \mathcal{X} \to \mathbb{R}$ is in this way obtained, with $1 \leqslant k < l \leqslant r$. These ranking functions can then be further used to generate multi-class probability estimates [48]. For a given data set, the ranking returned by each of the pairwise ranking functions is called bipartite, because it can be visualized by means of a bipartite graph in which the two subsets of nodes correspond to the data instances of the two classes and edges indicate the ranking order of two objects of different classes.

Ranking can be considered somewhere in the middle between pure discriminative modeling (we want good class predictions) and probability estimation (we want good predictions of class-conditional probabilities). The difference between both approaches is in the first place characterized by the type of loss or error function that is optimized. To this end, [1] introduced for ranking the concept of expected ranking accuracy as loss function. In a multi-class setting it can be formally introduced as follows.

Definition 2.1. Let $\mathcal{D}_j$ represent the conditional distribution over $\mathcal{X}$ given that the data object belongs to class $\mathcal{C}_j$ with $j = 1, \ldots, r$. For a set $\overline{\mathcal{F}} = \{f_{kl} \mid 1 \leqslant k < l \leqslant r\}$ of bipartite ranking functions, we define the pairwise expected ranking accuracy between classes $\mathcal{C}_k$ and $\mathcal{C}_l$ for the ranking function f_{kl} as

$$A_{kl}(f_{kl}) = \mathbf{Pr}_{\mathbf{X}_k \sim \mathcal{D}_k, \mathbf{X}_l \sim \mathcal{D}_l}\big\{f_{kl}(\mathbf{X}_k) < f_{kl}(\mathbf{X}_l)\big\} + \frac{1}{2}\mathbf{Pr}_{\mathbf{X}_k \sim \mathcal{D}_k, \mathbf{X}_l \sim \mathcal{D}_l}\big\{f_{kl}(\mathbf{X}_k) = f_{kl}(\mathbf{X}_l)\big\}. \tag{3}$$

For a single ranking function $f : \mathcal{X} \to \mathbb{R}$, the pairwise expected ranking accuracy is defined as

$$A_{kl}(f) = \mathbf{Pr}_{\mathbf{X}_k \sim \mathcal{D}_k, \mathbf{X}_l \sim \mathcal{D}_l}\big\{f(\mathbf{X}_k) < f(\mathbf{X}_l)\big\} + \frac{1}{2}\mathbf{Pr}_{\mathbf{X}_k \sim \mathcal{D}_k, \mathbf{X}_l \sim \mathcal{D}_l}\big\{f(\mathbf{X}_k) = f(\mathbf{X}_l)\big\}. \tag{4}$$

Here $\mathbf{X} \sim \mathcal{D}$ denotes that random vector $\mathbf{X}$ has distribution $\mathcal{D}$. Thus, the quality of the model is in essence evaluated by looking at the probability of correctly ranked couples $(\mathbf{X}_k, \mathbf{X}_l)$ of random vectors.[1] As in this definition, we will further always associate a single random vector $\mathbf{X}_j$ with each class, and without loss of generality, we may assume that these random vectors are independently sampled according to (different) unknown distributions, in which each distribution $\mathcal{D}_j$ corresponds to the data of one particular class. These unknown conditional distributions represent the probability of observing a certain input vector, given the class label of that input vector.

From a machine learning point of view, the primary concern is not to know the pairwise relationship of classes on a finite training set (represented by the empirical distribution, observed from a finite data sample). Rather, we want to find the relationship among the unknown underlying distributions $\mathcal{D}_j$, or in other words, the relationship between classes in input space. The r conditional class distributions $\mathcal{D}_j$, represented by random vectors $\mathbf{X}_j$, generate for each of the bipartite ranking functions f_{kl} two univariate distributions of prediction scores; for any two classes $\mathcal{C}_k$ and $\mathcal{C}_l$, two random variables $f_{kl}(\mathbf{X}_k)$ and $f_{kl}(\mathbf{X}_l)$ can be distinguished. In essence, we investigate whether the distributions $\mathcal{D}_j$ allow for an overall representation of these pairwise prediction score distributions as if they resulted from a single ranking function. Remark that the relationship between classes may not be interpreted here as a statistical dependence between classes, because data from different classes is of course independently sampled, and as such, the random vectors $\mathbf{X}_j$ are independent. We rather allude with the term relationship to the localization of the distributions in input space.

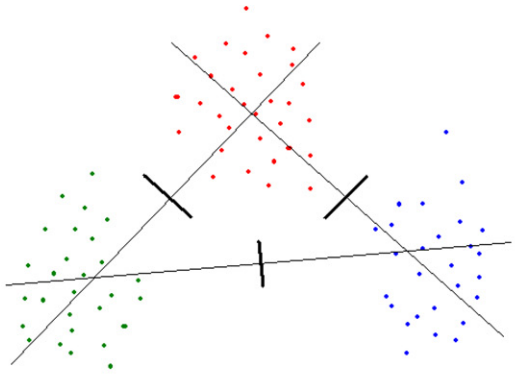
It is important to note that we will not require that the distributions of prediction scores generated by a single ranking function have to be identical to those generated by a set of bipartite ranking functions, since that would give too strong a condition. We will only enforce that the pairs of prediction score distributions have the same level of separability for both types of models, i.e. we require that the same pairwise expected ranking accuracies are obtained with a set of bipartite ranking functions and a single ranking function. The situation is graphically illustrated in Fig. 1 for a three-class classification problem. Three distributions (let's say $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$) of two-dimensional random vectors are shown (red, green, blue), together with two artificial triplets of pairwise output distributions having the same pairwise expected ranking accuracies: the triplet on top is generated from bipartite ranking functions and the triplet at the bottom from a single ranking function. One can easily verify that both models give rise to the same expected ranking accuracies, and because of that, this triplet of bipartite ranking functions will be called ERA ranking representable (see further).

For two classes $\mathcal{C}_k$ and $\mathcal{C}_l$, the expected ranking accuracy can be expressed in terms of the joint cumulative distribution function $F_{\mathbf{X}_k, \mathbf{X}_l}$ of the random vectors $\mathbf{X}_k$ and $\mathbf{X}_l$:

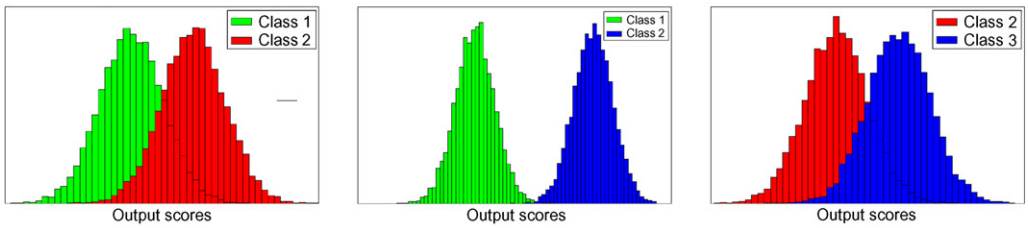
$$A_{kl}(f_{kl}) = \int\limits_{f_{kl}(\mathbf{x}_i) < f_{kl}(\mathbf{x}_j)} dF_{\mathbf{X}_k, \mathbf{X}_l}(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2} \int\limits_{f_{kl}(\mathbf{x}_i) = f_{kl}(\mathbf{x}_j)} dF_{\mathbf{X}_k, \mathbf{X}_l}(\mathbf{x}_i, \mathbf{x}_j).$$

As all random vectors are mutually independent, the joint cumulative distribution function of a couple can obviously be written as a product of its marginals.

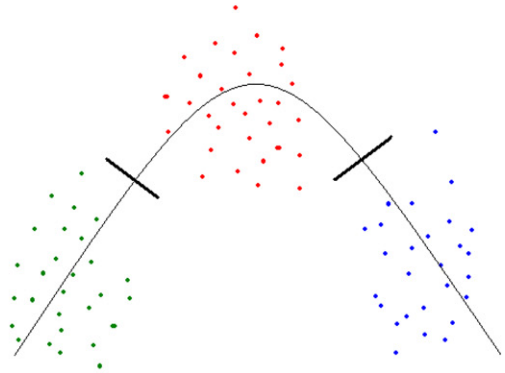
[1] For the remainder of our discussion, a restriction to vectorial input spaces is in fact not mandatory. We only make this restriction because random vector is a statistically more established concept than the more general *random data object*.



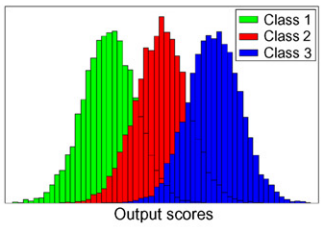
(a) Visualization of pairwise bipartite ranking functions in a three-class setting.



(b) Output score distributions obtained with bipartite ranking functions.



(c) Visualization of a single ranking function in a three-class setting.



(d) Output score distributions obtained with a single ranking function.

Fig. 1. A hypothetical example of pairwise output score distributions corresponding to an ERA ranking representable set of bipartite ranking functions (if the output score distributions originate from the unknown underlying distribution that generates the data) or an AUC ranking representable set of bipartite ranking functions (if the output score distributions originate from the observed empirical distribution on a finite data sample). The distributions obtained with a set of bipartite ranking functions are given on top, those obtained with a single ranking function are given at the bottom. For these distributions, both models generate the same triplets of bipartite ranking accuracies, because they have an identical level of separability in terms of ERA or AUC. See the electronic version of the paper for illustrations in color. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Proposition 2.2. *Given the extension $f_{lk} = -f_{kl}$, the pairwise expected ranking accuracies defined by* (3) *constitute the off-diagonal elements of a reciprocal relation. The same holds for the expected ranking accuracies defined by* (4)*.*

Proof. For a ranking function f_{kl}, let $F_{\mathbf{X}_k,\mathbf{X}_l}$ denote the joint cumulative distribution function for the random vectors $\mathbf{X}_k$ and $\mathbf{X}_l$. Similarly as in [16], we find

$$A_{kl}(f_{kl}) + A_{lk}(f_{lk}) = \int\limits_{f_{kl}(\mathbf{x}_i)<f_{kl}(\mathbf{x}_j)} dF_{\mathbf{X}_k,\mathbf{X}_l}(\mathbf{x}_i,\mathbf{x}_j) + \int\limits_{f_{kl}(\mathbf{x}_i)=f_{kl}(\mathbf{x}_j)} dF_{\mathbf{X}_k,\mathbf{X}_l}(\mathbf{x}_i,\mathbf{x}_j) + \int\limits_{f_{kl}(\mathbf{x}_i)>f_{kl}(\mathbf{x}_j)} dF_{\mathbf{X}_k,\mathbf{X}_l}(\mathbf{x}_i,\mathbf{x}_j) = 1.$$

Proving that (4) represents a reciprocal relation can be done in exactly the same way. □

Given the definition of expected ranking accuracy, we introduce the concept ERA ranking representability, as illustrated in Fig. 1.

Definition 2.3. Let $\mathbf{X}_1,\ldots,\mathbf{X}_r$ be r independent random vectors with respective conditional class distributions $\mathcal{D}_1,\ldots,\mathcal{D}_r$. We call a set $\overline{\mathcal{F}}$ of bipartite ranking functions ERA ranking representable on $\mathbf{X}_1,\ldots,\mathbf{X}_r$ if there exists a ranking function $f:\mathcal{X}\to\mathbb{R}$ such that for all $1\leqslant k<l\leqslant r$ it holds that

$$A_{kl}(f_{kl}) = A_{kl}(f). \tag{5}$$

The remainder of this article will be entirely dedicated to the quest for a way to verify ERA ranking representability. In essence, we are looking for a condition for which the set of bipartite ranking functions can be replaced by a single ranking function that gives evidence of the same expected ranking accuracy. We will see at the end that in that case the expected ranking accuracies satisfy a specific type of transitivity. This transitivity property will actually establish a condition on the distributions $\mathcal{D}_j$, but the condition itself will turn out to be distribution-independent, in the sense that the same condition must hold for any set of distributions $\mathcal{D}_1,\ldots,\mathcal{D}_r$. The details are given in Section 5, but we will first describe the finite sample case, for which some aspects of our story can be described in a less abstract way. Since the underlying distribution of the data is in general unknown, one obviously cannot compute the expected ranking accuracy, but one can estimate it on the basis of a finite labeled data sample $D=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$. This can be realized by computing the pairwise AUC, a nonparametric unbiased estimator of the expected ranking accuracy [1]. Thus, a ROC curve is constructed for each pair of classes. The AUC can be formally defined as follows [19,23,40].

Definition 2.4. For a set $\overline{\mathcal{F}}$ of bipartite ranking functions, we define the pairwise AUC between classes $\mathcal{C}_k$ and $\mathcal{C}_l$ for the ranking function f_{kl} with $1\leqslant k<l\leqslant r$ as

$$\widehat{A}_{kl}(f_{kl},D) = \frac{1}{n_k n_l}\sum_{y_i=\mathcal{C}_k}\sum_{y_j=\mathcal{C}_l} I_{f_{kl}(\mathbf{x}_i)<f_{kl}(\mathbf{x}_j)}. \tag{6}$$

For a single ranking function $f:\mathcal{X}\to\mathbb{R}$, the pairwise AUC is defined as

$$\widehat{A}_{kl}(f,D) = \frac{1}{n_k n_l}\sum_{y_i=\mathcal{C}_k}\sum_{y_j=\mathcal{C}_l} I_{f(\mathbf{x}_i)<f(\mathbf{x}_j)}.$$

Remark that I denotes the indicator function that returns one when its argument is true and zero otherwise.

For further details on this definition and a general discussion of ROC analysis in multi-class settings, we refer for example to [20,21,24,28]. Interestingly, it has been shown by [8,29,49] that the binary AUC is equivalent to the *Wilcoxon–Mann–Whitney* statistic. It measures the expected ranking accuracy on the empirical distribution instead of the unknown underlying distribution and, by definition, it also satisfies the reciprocity property. Given a finite data sample, the AUC allows us to define the following form of ranking representability that can be interpreted as ERA ranking representability of the observed empirical distribution.

Definition 2.5. We call a set $\overline{\mathcal{F}}$ of bipartite ranking functions AUC ranking representable on D if there exists a ranking function $f:\mathcal{X}\to\mathbb{R}$ such that for all $1\leqslant k<l\leqslant r$ it holds that

$$\widehat{A}_{kl}(f_{kl},D) = \widehat{A}_{kl}(f,D). \tag{7}$$

In [46], we introduced AUC ranking representability as a relaxation of strict ranking representability, which basically assumes that all bipartite ranking functions must be consistent with a global ranking function. We showed that strict ranking representability can be easily verified by investigating whether a graph is free of cycles. Unfortunately, strict ranking representability has a very limited applicability, since it is a condition that cannot be satisfied for realistic data samples.

However, from a statistical perspective, such a strong condition is not required and that was our main motivation to relax this condition to AUC ranking representability.

AUC ranking representability can be illustrated too by Fig. 1, but now the univariate output score distributions are generated from empirical multivariate distributions. Consequently, AUC ranking representability can be easily verified for small data samples by enumerating all possible rankings of the data and computing for each of them the pairwise AUCs, as shown by the example in the introduction.

3. Transitivity

In this section, we give a detailed introduction to the framework of cycle transitivity [11], which has quite recently been put forward as a unification of fuzzy transitivity on the one hand and stochastic transitivity [37,45] on the other hand. In [10] it was shown that cycle transitivity covers *FG*-transitivity, a slightly older unifying framework for fuzzy and stochastic transitivity. Moreover, other types than fuzzy or stochastic transitivity can be elegantly expressed in the cycle transitivity framework. We will give a brief overview of some types of cycle transitivity that are relevant for our discussion.

3.1. Notations

Let $Q : \mathcal{X}^2 \to [0,1]$ be a reciprocal relation defined on a set of data objects $\mathcal{X}$. For any $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{X}^2$, we first introduce the shorthand notation $Q_{ij} = Q(\mathbf{x}_i, \mathbf{x}_j)$. For any $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \in \mathcal{X}^3$ we define

$$\alpha_{123} = \min(Q_{12}, Q_{23}, Q_{31}), \qquad \beta_{123} = \text{median}(Q_{12}, Q_{23}, Q_{31}), \qquad \gamma_{123} = \max(Q_{12}, Q_{23}, Q_{31}).$$

3.2. Product transitivity

Product transitivity (further denoted $T_{\mathbf{P}}$-transitivity) can be considered as a specific type of T-transitivity, a popular notion in fuzzy set theory. Further we will give a formal definition of T-transitivity, but here we briefly mention that the product t-norm $T_{\mathbf{P}}(a,b) = ab$ gives rise to a type of T-transitivity and it also forms the basis for the introduction of cycle transitivity, as shown in [11]. A reciprocal relation satisfies product transitivity if for any $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \in \mathcal{X}^3$ it holds that

$$Q_{12} Q_{23} \leqslant Q_{13}. \tag{8}$$

Let us now consider a single triplet $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$. When all permutations of this triplet are considered and the reciprocity property is taken into account, then (8) gives rise to the following six conditions on Q_{12}, Q_{23}, Q_{31}:

$$\begin{aligned} &Q_{12} Q_{23} \leqslant Q_{13}, \qquad Q_{13} Q_{32} \leqslant Q_{12}, \\ &Q_{23} Q_{31} \leqslant Q_{21}, \qquad Q_{21} Q_{13} \leqslant Q_{23}, \\ &Q_{31} Q_{12} \leqslant Q_{32}, \qquad Q_{32} Q_{21} \leqslant Q_{31}. \end{aligned}$$

De Baets et al. [11] showed that these six inequalities can be reduced to one double inequality, expressed in terms of α, β, γ:

$$\beta_{123}\gamma_{123} \leqslant \alpha_{123} + \beta_{123} + \gamma_{123} - 1 \leqslant 1 - (1-\alpha_{123})(1-\beta_{123}). \tag{9}$$

If $L(\beta_{123}, \gamma_{123})$ and $U(\alpha_{123}, \beta_{123})$ represent the lower and upper bound in the above expression, then the following identity between both bounds is observed:

$$L(\beta_{123}, \gamma_{123}) = 1 - U(1-\gamma_{123}, 1-\beta_{123}).$$

Moreover, the obtained lower and upper bound are indifferent to any permutation of $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ and the double inequality holds for both directions of the loop.

3.3. Definition of cycle transitivity

The observation made to rewrite $T_{\mathbf{P}}$-transitivity as the double inequality (9) lays the foundation of cycle transitivity. Within the framework of cycle transitivity, the upper bound (and corresponding lower bound) are generalized towards other bounds than the ones given above. To this end, let us define $\Delta = \{(\alpha, \beta, \gamma) \in [0,1]^3 \mid \alpha \leqslant \beta \leqslant \gamma\}$ and consider a function $U : \Delta \to \mathbb{R}$, then, by analogy with (9), we can call a reciprocal preference relation $Q : \mathcal{X}^2 \to [0,1]$ cycle-transitive w.r.t. U if for any $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \in \mathcal{X}^3$ it holds that

$$1 - U(1-\gamma_{123}, 1-\beta_{123}, 1-\alpha_{123}) \leqslant \alpha_{123} + \beta_{123} + \gamma_{123} - 1 \leqslant U(\alpha_{123}, \beta_{123}, \gamma_{123}).$$

Contrary to the derivation above for $T_{\mathbf{P}}$-transitivity, the upper bound function U takes in general three arguments instead of two. In case of $T_{\mathbf{P}}$-transitivity, the upper bound function becomes

$$U_{T_{\mathbf{P}}}(\alpha, \beta, \gamma) = \alpha + \beta - \alpha\beta.$$

The case where $U(\alpha,\beta,\gamma)=\beta$ turns out to be another form of fuzzy transitivity as discussed in Section 3.4. The double inequality leads to two conditions: the lower bound should not exceed the upper bound and the value $\alpha+\beta+\gamma-1$ should be located between both bounds.

Definition 3.1. A function $U:\Delta\to\mathbb{R}$ is called an upper bound function if it satisfies the following properties:

(1) $U(0,0,1)\geqslant 0$ and $U(0,1,1)\geqslant 1$,
(2) for any $\alpha,\beta,\gamma\in\Delta$:

$$U(\alpha,\beta,\gamma)+U(1-\gamma,1-\beta,1-\alpha)\geqslant 1. \tag{10}$$

The definition of an upper bound function does not include any monotonicity condition. We define the dual lower bound function $L:\Delta\to\mathbb{R}$ of a given upper bound function U as

$$L(\alpha,\beta,\gamma)=1-U(1-\gamma,1-\beta,1-\alpha),$$

implying that $L\leqslant U$ when (10) holds. These tools allow us to define formally the notion of cycle transitivity.

Definition 3.2. A reciprocal relation $Q:\mathcal{X}^2\to[0,1]$ is called cycle-transitive w.r.t. an upper bound function U if for any $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)\in\mathcal{X}^3$ it holds that

$$L(\alpha_{123},\beta_{123},\gamma_{123})\leqslant\alpha_{123}+\beta_{123}+\gamma_{123}-1\leqslant U(\alpha_{123},\beta_{123},\gamma_{123}), \tag{11}$$

where L is the dual lower bound function of U.

From this construction immediately follows that, as soon as the double inequality is fulfilled for a triplet $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)\in\mathcal{X}^3$, it is also fulfilled for any permutation of the triplet. Therefore, in practice one only needs to check (11) for a single permutation of $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$. Alternatively, due to the same duality, one can also opt to verify only the upper bound, or equivalently the lower bound, for two permutations of $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$ that are not cyclic permutations of one another. This is summarized as follows.

Proposition 3.3. *(See [11].) A reciprocal relation $Q:\mathcal{X}^2\to[0,1]$ is cycle-transitive w.r.t. an upper bound function U if for any $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)\in\mathcal{X}^3$ it holds that*

$$\alpha_{123}+\beta_{123}+\gamma_{123}-1\leqslant U(\alpha_{123},\beta_{123},\gamma_{123}). \tag{12}$$

The loosest upper bound one can choose is the constant function $U=2$, which means that there is no restriction on the values the reciprocal relation can take. It will become clear later that the upper bound function represents a very straightforward way to link different types of transitivity and, in particular, to determine whether a particular form of transitivity follows from another form of transitivity. For example, given two types of transitivity A and B that can be casted in the framework of cycle transitivity by means of upper bound functions U_A and U_B such that $U_A(\alpha,\beta,\gamma)\leqslant U_B(\alpha,\beta,\gamma)$, we automatically know that type-A transitivity implies type-B transitivity. It is shown in the following sections that the cycle transitivity framework incorporates various types of transitivity.

3.4. Fuzzy transitivity

T-transitivity is an important notion in the fuzzy set literature and it is a desirable property of fuzzy relations. In this work we will only consider the case where fuzzy relations are reciprocal relations, a condition that does not hold in general. The traditional definition given in terms of t-norms can be generalized to the more general class of conjunctors. As shown by [11], we will start with this more general case in order to establish the link with cycle transitivity.

Definition 3.4. A binary operation $C:[0,1]^2\to[0,1]$ is called a conjunctor if it satisfies the following properties:

(1) Its restriction to $\{0,1\}^2$ coincides with the boolean conjunction.
(2) Monotonicity: C is increasing in both variables.

This gives us the opportunity to define C-transitivity.

Definition 3.5. Let C be a conjunctor. A fuzzy relation $R:\mathcal{X}^2\to[0,1]$ is called C-transitive if for any $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)\in\mathcal{X}^3$ it holds that

$$C\big(R(\mathbf{x}_1,\mathbf{x}_2),R(\mathbf{x}_2,\mathbf{x}_3)\big)\leqslant R(\mathbf{x}_1,\mathbf{x}_3). \tag{13}$$

Two important families of conjunctors are t-norms (neutral element 1, monotonicity, commutativity, associativity) and copulas (neutral element 1, absorbing element 0, monotonicity, 2-increasingness). The 2-increasingness property is relaxed to the 1-Lipschitz condition for quasi-copulas. A close relationship exists between copulas and t-norms, since t-norms having the 1-Lipschitz property correspond to associative copulas. Three important t-norms (and copulas) that will appear further in this article are the minimum t-norm $T_{\mathbf{M}}(a,b) = \min(a,b)$, the product t-norm $T_{\mathbf{P}}(a,b) = ab$ and the Łukasiewicz t-norm $T_{\mathbf{L}}(a,b) = \max(a+b-1,0)$. Given the restriction to reciprocal relations, the following proposition characterizes the reformulation of C-transitivity in terms of cycle transitivity.

Proposition 3.6. *(See [11].) Let C be a commutative conjunctor such that $C \leqslant T_{\mathbf{M}}$. A reciprocal relation $Q : \mathcal{X}^2 \to [0,1]$ is C-transitive if and only if it is cycle-transitive w.r.t. the upper bound function U_C defined by*

$$U_C(\alpha,\beta,\gamma) = \min\big(\alpha+\beta-C(\alpha,\beta), \alpha+\gamma-C(\alpha,\gamma), \beta+\gamma-C(\beta,\gamma)\big).$$

If C is 1-Lipschitz, then the upper bound function can be simplified to

$$U_C(\alpha,\beta,\gamma) = \alpha+\beta-C(\alpha,\beta).$$

The three t-norms discussed above define the following upper bound functions:

$$U_{T_{\mathbf{M}}}(\alpha,\beta,\gamma) = \beta, \qquad U_{T_{\mathbf{P}}}(\alpha,\beta,\gamma) = \alpha+\beta-\alpha\beta, \qquad U_{T_{\mathbf{L}}}(\alpha,\beta,\gamma) = 1.$$

As shown in [46, submitted], $T_{\mathbf{L}}$-transitivity is equivalent to the triangle inequality. We remark that the triangle inequality is traditionally used for symmetric relations, but it has been considered too by [37] and [36] as a property to characterize reciprocal preference relations.

3.5. Stochastic transitivity

We first introduced fuzzy transitivity for its straightforward reformulation in terms of cycle transitivity. On the other hand, stochastic transitivity is a fairly different framework for characterizing reciprocal relations. Historically, it has played a more dominant role than fuzzy transitivity. As mentioned in the introduction, stochastic transitivity is closely connected to ranking representability of reciprocal relations.

Definition 3.7. Let g be an increasing $[1/2,1]^2 \to [0,1]$ mapping. A reciprocal relation $Q : \mathcal{X}^2 \to [0,1]$ is called g-stochastically transitive if for any $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) \in \mathcal{X}^3$ it holds that

$$\big(Q(\mathbf{x}_1,\mathbf{x}_2) \geqslant 1/2 \wedge Q(\mathbf{x}_2,\mathbf{x}_3) \geqslant 1/2\big) \quad \Rightarrow \quad Q(\mathbf{x}_1,\mathbf{x}_3) \geqslant g\big(Q(\mathbf{x}_1,\mathbf{x}_2), Q(\mathbf{x}_2,\mathbf{x}_3)\big).$$

Many specific types of stochastic transitivity can be found in the literature, such as weak stochastic transitivity ($g(a,b) = 1/2$), moderate stochastic transitivity ($g(a,b) = \min(a,b)$) and strong stochastic transitivity ($g(a,b) = \max(a,b)$).

3.6. Product-based upper bound functions

Another class of interesting upper bound functions is inspired by the upper bound of product transitivity. More specifically, the following three upper bound functions are closely related:

(1) strong product transitivity: $U_{T_{\mathbf{P}}}(\alpha,\beta,\gamma) = \alpha+\beta-\alpha\beta$,
(2) moderate product transitivity: $U_{mp}(\alpha,\beta,\gamma) = \alpha+\gamma-\alpha\gamma$,
(3) weak product transitivity: $U_D(\alpha,\beta,\gamma) = \beta+\gamma-\beta\gamma$.

The first one characterizes the traditional transitivity w.r.t. the product t-norm, as explained before. The second one has recently found an application in the field of partially ordered sets [13]. This upper bound function will reappear at the end of this paper. The third upper bound function characterizes reciprocal relations that are generated by collections of dice [15]. Because of that, this type of transitivity has been called *dice transitivity*, and we will need it to analyze the ranking representability of bipartite ranking functions. We can also mention for the sake of completeness that dice transitivity so far found an application in the comparison of independent random variables [16]. Other upper bound functions (based on $T_{\mathbf{L}}$ or $T_{\mathbf{M}}$) can be linked to the comonotone or counter-monotone comparison of random variables [14,17].

4. AUC ranking representability

In order to analyze the relationship between AUC and ERA ranking representability on the one hand and cycle transitivity on the other hand, we first rephrase some important results that have been obtained recently for reciprocal relations that are generated from collections of dice, in a game-theoretic context [15]. We will for simplicity omit here the discussion on

the relationship between pairwise AUCs and the reciprocal relations in this game-theoretic setting, but one can easily show that both types of relations exhibit the same transitivity conditions as necessary conditions. For more details we refer to [46], in which it was shown that an AUC ranking representable set $\overline{\mathcal{F}}$ possesses a specific form of cycle transitivity, namely dice transitivity.

Proposition 4.1. *(See [15].) The reciprocal relation of pairwise AUCs generated by an AUC ranking representable set $\overline{\mathcal{F}}$ of bipartite ranking functions is dice-transitive.*

It is important to note that the data set must be identically and independently distributed for this proposition, as often assumed in machine learning, and mentioned above.

As indicated in the previous section, dice transitivity is stronger than the triangle inequality. Subsequently, the above important result was extended in [16] for independent random variables. In a nutshell, the same transitivity condition applies when independent random variables are compared in a pairwise way, leading to a reciprocal relation that closely resembles the expected ranking accuracy. The simplest way to see this correspondence is to observe that for independent random vectors $\mathbf{X}$ and $\mathbf{X}'$, the output scores $f(\mathbf{X})$ and $f(\mathbf{X}')$ are independent random variables as well. Reformulated in the current discussion, this yields the following important result.

Proposition 4.2. *(See [16].) Let $\mathbf{X}_1, \ldots, \mathbf{X}_r$ be r independent random vectors. The reciprocal relation of pairwise expected ranking accuracies generated by an ERA ranking representable set $\overline{\mathcal{F}}$ of bipartite ranking functions is dice-transitive.*

From the first proposition we know that the pairwise AUCs generated from an AUC ranking representable set of bipartite ranking functions give evidence of a particular form of transitivity, stronger than the triangle inequality, but weaker than $T_{\mathbf{P}}$-transitivity. De Schuymer et al. [15] gave counterexamples to illustrate that the reciprocal relation of pairwise AUCs generated from an AUC ranking representable set of bipartite ranking functions not always satisfies $T_{\mathbf{P}}$-transitivity.

So, dice transitivity gives rise to a necessary condition for AUC ranking representability, but is it also a sufficient condition? The answer is definitely negative, since even much stronger types of transitivity not necessarily lead to AUC ranking representability. We showed for example that strong stochastic transitivity and even $T_{\mathbf{M}}$-transitivity are not sufficient for AUC ranking representability. However, in Section 5 it will follow from our results that $T_{\mathbf{M}}$-transitivity becomes sufficient for ERA ranking representability. In order to describe a sufficient condition, we have to introduce a graph-theoretic reformulation of AUC ranking representability.

Definition 4.3. Let $\overline{\mathcal{F}}$ be a set of bipartite ranking functions. We define $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ as the set of complete directed graphs $G = (V, E)$ with V the set of nodes and E the set of edges, so that the following three properties hold:

(1) Each node v_i in V is associated with one data object $(\mathbf{x}_i, y_i)$ in D.
(2) No cycles occur in the subsets $V_k = \{v_i \in V \mid y_i = \mathcal{C}_k\}$.
(3) For $1 \leqslant k < l \leqslant r$:

$$\widehat{A}_{kl}(f_{kl}, D) = \frac{|\{(v_i, v_j) \in E \mid y_i = \mathcal{C}_k \wedge y_j = \mathcal{C}_l\}|}{n_k n_l}. \tag{14}$$

Let us try to express this definition in a less formal way. We are in essence looking for all graphs $G = (V, E)$ in which we associate one data object from the data set with a node such that we obtain r subsets $V_1, \ldots, V_r$ for r classes. We require in addition that the nodes within each subset are ordered (which results in an acyclic subgraph for these subsets), and that the fraction of edges from subset V_k to V_l corresponds to $\widehat{A}_{kl}(f_{kl}, D)$. Some examples of such graphs will be presented later on. In this way, we only consider complete directed graphs. Remark that a complete directed graph is a graph in which each pair of nodes is connected by exactly one (directed) edge. So, $(v, v') \in E$ implies $(v', v) \notin E$.

It follows directly from the definition that $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ cannot be empty. Its cardinality will usually be greater than 1 since different graphs satisfying (14) will be found for a given $\overline{\mathcal{F}}$ and D. In the following proposition, AUC ranking representability is reformulated in terms of these graphs.

Definition 4.4. We introduce $\mathfrak{H}_{AUC}(\overline{\mathcal{F}}, D)$ as the subset of $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ containing only directed acyclic graphs (DAGs).

Proposition 4.5. *A set $\overline{\mathcal{F}}$ of bipartite ranking functions is AUC ranking representable on D if and only if $\mathfrak{H}_{AUC}(\overline{\mathcal{F}}, D)$ is not empty.*

Using the graph-theoretic concepts introduced above, we have a sufficient condition for AUC ranking representability. Nevertheless, this condition cannot be verified for large datasets, since the cardinality of $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ exponentially increases with the size of D. Similar to [15], we further examine the three-class case in order to find a sufficient condition that can be verified more easily. The reason for this restriction is that we will need cycle transitivity (which has so far only been defined on triplets). The results obtained for three classes can then be further extended to more classes with approximation techniques. We start with introducing a new type of transitivity.

Definition 4.6. An (a^*, s)-split, denoted $\mathbf{a}^*$, is an increasing ordered list (or vector) $\mathbf{a}^* = (a_1^*, a_2^*, \ldots, a_s^*)$ of s (not necessarily strictly) positive integers summing up to a^*. An (a^*, s, t)-split is an (a^*, s)-split for which each component of $\mathbf{a}^*$ is upper bounded by t. The set of all (a^*, s, t)-splits will be denoted $\mathfrak{S}(a^*, s, t)$. We define the dual $\mathbf{b}$ of an (a^*, s, t)-split as the decreasing vector $\mathbf{b}^* = (a_s^*, a_{s-1}^*, \ldots, a_1^*)$. The set of all dual (a^*, s, t)-splits will be denoted $\widetilde{\mathfrak{S}}(a^*, s, t)$.

Example 4.7. We give two simple examples to illustrate the above definition:

$$\mathfrak{S}(10,4,3) = \big\{(1,3,3,3),(2,2,3,3)\big\},$$
$$\widetilde{\mathfrak{S}}(11,3,6) = \big\{(6,5,0),(6,4,1),(6,3,2),(5,5,1),(5,4,2),(5,3,3)\big\}.$$

Definition 4.8. Let $(n_1, \ldots, n_r) \in \mathbb{N}^r$ and let

$$\mho_{kl} = \left\{ a \in [0,1] \;\middle|\; (\exists a^* \in \mathbb{N})\left(a = \frac{a^*}{n_k n_l}\right) \right\}.$$

The family of functions $C_{jkl} : \mho_{jk} \times \mho_{kl} \to \mho_{jl}$ is defined by:

$$C_{jkl}(a,b) = \frac{1}{n_j n_l} \min_{\substack{\mathbf{a}^* \in \mathfrak{S}(a^*, n_k, n_j) \\ \mathbf{b}^* \in \widetilde{\mathfrak{S}}(b^*, n_k, n_l)}} \sum_{i=1}^{n_k} \big(a_i^* - a_{i-1}^*\big) b_i^*,$$

for $j, k, l \in \{1, \ldots, r\}$.

The value $C_{jkl}(a,b)$ is the solution of an integer quadratic program. To illustrate this, let us rewrite the minimization as:

$$\min_{\mathbf{a}^*, \mathbf{b}^*} \frac{1}{n_j n_l} \sum_{i=1}^{n_k} \big(a_i^* - a_{i-1}^*\big) b_i^*$$

$$\text{subject to } \begin{cases} \sum_{i=1}^{n_k} a_i^* = a^*, \\ \sum_{i=1}^{n_k} b_i^* = b^*, \\ a_i^* \geqslant a_{i-1}^*, \quad \forall i \in \{1, \ldots, n_k\}, \\ b_i^* \leqslant b_{i-1}^*, \quad \forall i \in \{2, \ldots, n_k + 1\}, \\ 0 \leqslant a_i^* \leqslant n_j, \quad \forall i \in \{1, \ldots, n_k\}, \\ 0 \leqslant b_i^* \leqslant n_l, \quad \forall i \in \{1, \ldots, n_k\}, \\ a_i^*, b_i^* \in \mathbb{N}, \quad \forall i \in \{1, \ldots, n_k\}, \\ a_0^* = 0, \quad b_{n_k+1}^* = 0. \end{cases} \tag{15}$$

In Fig. 2 the family of functions C_{jkl} is visualized for some (small) n_j, n_k and n_l. The function values were computed by exhaustively verifying all feasible solutions of the integer quadratic program, which can only be done for small values of n_j, n_k and n_l.

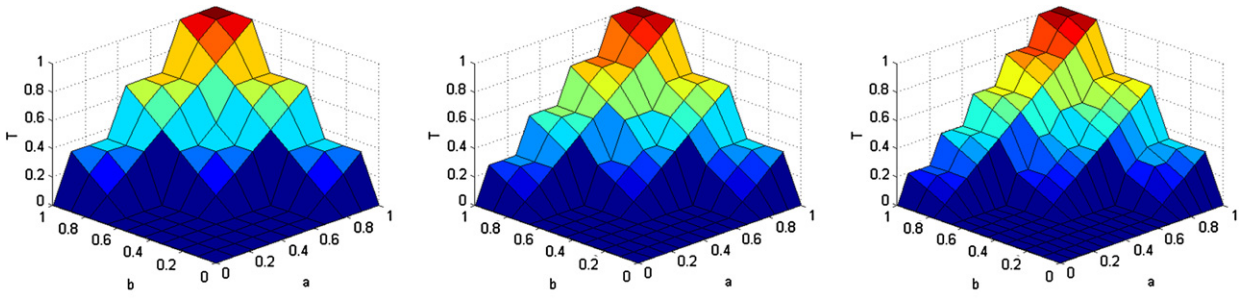
Example 4.9. Let us consider the situation: $n_j = 3$, $n_k = 4$, $n_l = 5$, $a = 9/12$, $b = 13/20$. The objective is minimized over the following splits:

$$\mathfrak{S}(9,4,3) = \big\{(0,3,3,3),(1,2,3,3),(2,2,2,3)\big\},$$
$$\widetilde{\mathfrak{S}}(13,4,5) = \big\{(5,5,3,0),(5,5,2,1),(5,4,4,0),(5,4,3,1),(5,4,2,2),(5,3,3,2),$$
$$(4,4,4,1),(4,4,3,2),(4,3,3,3)\big\}.$$

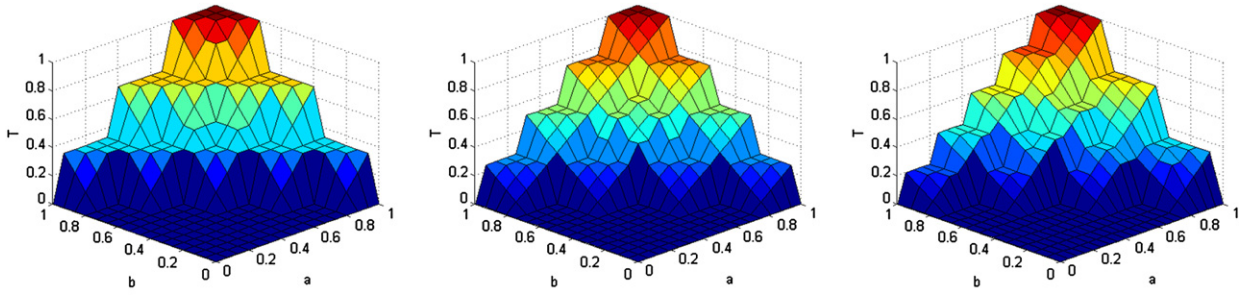
The minimum of the objective function is obtained for the splits $\mathbf{a}^* = (2,2,2,3)$ and $\mathbf{b}^* = (4,3,3,3)$ such that $C_{jkl}(a,b) = 11/15$. It turns out that for this example (and many other cases) the minimum can be found without computing the objective function for all splits exhaustively. To understand this, we have to reveal the graph-theoretic interpretation from which the integer quadratic program originates.

Proposition 4.10. *Given the graph-theoretic reformulation of Definition* 4.3*, we have*

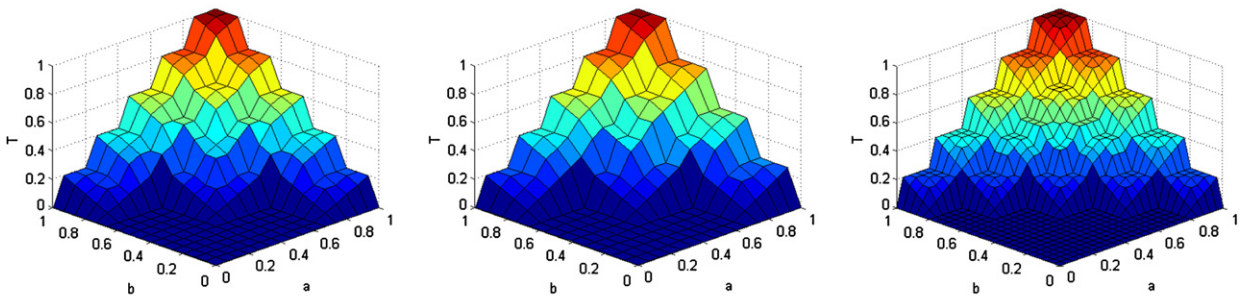
$$C_{jkl}(\widehat{A}_{jk}, \widehat{A}_{kl}) = \min_{G \in \mathfrak{H}_{AUC}(\overline{\mathcal{F}}, D)} \frac{|\{(v_a, v_c) \in V_j \times V_l \mid (\exists v_b \in V_k)((v_a, v_b), (v_b, v_c) \in E)\}|}{n_j n_l}.$$



(a) $n_j = 3, n_k = 3, n_l = 3$ (b) $n_j = 3, n_k = 3, n_l = 4$ (c) $n_j = 3, n_k = 3, n_l = 5$



(d) $n_j = 3, n_k = 5, n_l = 3$ (e) $n_j = 4, n_k = 4, n_l = 4$ (f) $n_j = 3, n_k = 4, n_l = 5$



(g) $n_j = 5, n_k = 3, n_l = 5$ (h) $n_j = 4, n_k = 3, n_l = 5$ (i) $n_j = 5, n_k = 5, n_l = 5$

Fig. 2. The family of functions C_{jkl} visualized.

The proof has been given in [46]. Nonetheless, let us briefly explain in words what this equation means, since the graph-theoretic interpretation leads to a crucial insight for a further understanding of this paper. Given a data set D and a set of bipartite ranking functions $\overline{\mathcal{F}}$, we examine all couples of nodes $(v_a, v_c) \in V_j \times V_l$ in all graphs $G \in \mathfrak{H}_{AUC}(\overline{\mathcal{F}}, D)$. The proposition states that C_{jkl} equals the minimal number of such couples connected by a path passing through a node of layer V_k over all these graphs. For any graph $G \in \mathfrak{H}_{AUC}(\overline{\mathcal{F}}, D)$, we have that a^* represents the number of edges departing from a node v_a of subset V_j and ending in a node v_b of subset V_k. Similarly, b^* represents the number of edges departing from a node v_b of subset V_k and arriving in a node v_c of subset V_l.

In Section 5, examples will be provided to illustrate the graph-theoretic interpretation of the integer quadratic program (see Figs. 6, 7 and 8). Based on this graph-theoretic interpretation, let us introduce a new type of transitivity.

Definition 4.11. A reciprocal relation of pairwise AUCs $\widehat{A}_{kl}(f_{kl}, D)$ is called AUC transitive if for all $j, k, l \in \{1, \ldots, r\}$ it holds that

$$C_{jkl}(\widehat{A}_{jk}, \widehat{A}_{kl}) \leqslant \widehat{A}_{jl}. \tag{16}$$

We emphasize that this type of transitivity in certain sense differs from all existing types of transitivity, since the condition that a given triplet of values must satisfy depends on their indices.

Proposition 4.12. *Let $\overline{\mathcal{F}} = \{f_{12}, f_{23}, f_{13}\}$ be a triplet of bipartite ranking functions. The corresponding triplet of pairwise AUCs is AUC transitive on D if and only if $\mathfrak{H}_{AUC}(\overline{\mathcal{F}}, D)$ is not empty.*

Corollary 4.13. *A triplet $\overline{\mathcal{F}} = \{f_{12}, f_{23}, f_{13}\}$ of bipartite ranking functions is AUC ranking representable on D if and only if the corresponding reciprocal relation of AUCs is AUC transitive.*

Corollary 4.14. *AUC transitivity implies dice transitivity.*

Briefly summarized, we have in essence proven that AUC transitivity implies dice transitivity, by using a chain of equivalences and implications. Firstly, we observed in Proposition 4.12 that the AUC transitivity of pairwise AUCs corresponds to the existence of a DAG in the graph set $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$. Secondly, by applying Proposition 4.5, the existence of a DAG in $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ allowed us to conclude that the corresponding set $\overline{\mathcal{F}}$ of bipartite ranking functions is AUC ranking representable. Thirdly, due to the AUC ranking representability of $\overline{\mathcal{F}}$, we were able to express the pairwise AUCs as a reciprocal relation originating from collections of dice. Finally, from Proposition 4.1 it followed that this reciprocal relation is dice-transitive, so that AUC transitivity implies dice transitivity.

Given the results obtained in previous work, we are able to prove in addition a number of new interesting properties for the family of functions C_{jkl}. In the next section, where the generalization to ERA ranking representability is described, these properties will reduce to some well-known characteristics of t-norms.

Proposition 4.15. *Let $(n_1, \ldots, n_r) \in \mathbb{N}^r$. The family of functions $C_{jkl} : \mathfrak{V}_{jk} \times \mathfrak{V}_{kl} \to \mathfrak{V}_{jl}$ as in Definition 4.8 has the following properties*:

(1) $\forall j,k,l \in \{1,\ldots,r\}$: *$C_{jkl}$ is increasing in both variables.*
(2) $\forall j,k,l \in \{1,\ldots,r\}, \forall (a,b) \in \mathfrak{V}_{jk} \times \mathfrak{V}_{kl}$: $C_{jkl}(a,b) = C_{lkj}(b,a)$.
(3) $\forall j,k,l \in \{1,\ldots,r\}, \forall a \in \mathfrak{V}_{jk}$: $C_{jkl}(a,0) = 0$.
(4) $\forall j,k,l \in \{1,\ldots,r\}, \forall a \in \mathfrak{V}_{jk}$: $C_{jkl}(a,1) = \frac{1}{n_j}\lceil \frac{a^*}{n_k} \rceil$.
(5) $\forall j,k,l \in \{1,\ldots,r\}, \forall (a,b) \in \mathfrak{V}_{jk} \times \mathfrak{V}_{kl}$:

$$C_{jkl}(a,b) \leqslant \frac{1}{n_j n_l} \left\lceil \frac{a^*}{n_k} \right\rceil \left\lceil \frac{b^*}{n_k} \right\rceil.$$

(6) $\forall j,k,l \in \{1,\ldots,r\}, \forall (a,b) \in \mathfrak{V}_{jk} \times \mathfrak{V}_{kl}$:

$$n_j a \in \mathbb{N} \wedge n_l b \in \mathbb{N} \quad \Rightarrow \quad C_{jkl}(a,b) \leqslant ab,$$

with $a = a^*/(n_j n_k)$, $b = b^*/(n_k n_l)$ and $\lceil \cdot \rceil : \mathbb{R} \to \mathbb{N}$ the ceiling function that retrieves the closest integer greater than or equal to a given real number.

Proof.

Property 1. The objective function of optimization problem (15) is an increasing function of a^* and b^*. This property directly follows from the graph-theoretic interpretation of the integer quadratic program. When a^* or b^* increases, then respectively, the number of incoming or outgoing edges in the layer V_k increases. As a consequence, also the minimum of the objective function increases, since the minimum corresponds to the number of connected couples from $V_j \times V_l$, connected via a node in V_k.

Property 2. Let us define $b_{n_k+1} = 0$. We find:

$$C_{jkl}(a,b) = \frac{1}{n_j n_l} \min_{\substack{\mathbf{a}^* \in \mathfrak{S}(a^*, n_k, n_j) \\ \mathbf{b}^* \in \widetilde{\mathfrak{S}}(b^*, n_k, n_l)}} \sum_{i=1}^{n_k} a_i^* b_i^* - \sum_{i=1}^{n_k} a_{i-1}^* b_i^* = \frac{1}{n_j n_l} \min_{\substack{\mathbf{a}^* \in \mathfrak{S}(a^*, n_k, n_j) \\ \mathbf{b}^* \in \widetilde{\mathfrak{S}}(b^*, n_k, n_l)}} \sum_{i=1}^{n_k} a_i^* b_i^* - \sum_{i=1}^{n_k} a_i^* b_{i+1}^*$$

$$= \frac{1}{n_j n_l} \min_{\substack{\mathbf{a}^* \in \mathfrak{S}(a^*, n_k, n_j) \\ \mathbf{b}^* \in \widetilde{\mathfrak{S}}(b^*, n_k, n_l)}} \sum_{i=1}^{n_k} a_i^* \left(b_i^* - b_{i+1}^*\right) = \frac{1}{n_j n_l} \min_{\substack{\mathbf{a}^* \in \widetilde{\mathfrak{S}}(a^*, n_k, n_j) \\ \mathbf{b}^* \in \mathfrak{S}(b^*, n_k, n_l)}} \sum_{i=1}^{n_k} a_i^* \left(b_i^* - b_{i-1}^*\right) = C_{lkj}(b,a).$$

Property 3. When we fill in 0 for a^* or b^* in optimization problem (15), then the solution of the integer quadratic program is 0.

Property 4. When $b = 1$, then $b_i^* = n_l$ for all $i \in \{1,\ldots,n_j\}$. Taking into account that $a = a^*/(n_j n_k)$, we find:

$$C_{jkl}(a,1) = \frac{1}{n_j n_l} \min_{\substack{\mathbf{a}^* \in \mathfrak{S}(a^*, n_k, n_j) \\ \mathbf{b}^* \in \widetilde{\mathfrak{S}}(b^*, n_k, n_l)}} \sum_{i=1}^{n_k} \left(a_i^* - a_{i-1}^*\right) n_l = \frac{1}{n_j} \min_{\mathbf{a}^* \in \mathfrak{S}(a^*, n_k, n_j)} \sum_{i=1}^{n_k} \left(a_i^* - a_{i-1}^*\right) = \frac{1}{n_j} \min_{\mathbf{a}^* \in \mathfrak{S}(a^*, n_j, n_j)} a_{n_k}^*$$
$$= \frac{1}{n_j} \left\lceil \frac{a^*}{n_k} \right\rceil.$$

Property 5. To understand this property, again the graph-theoretic interpretation of $C_{jkl}(a,b)$ as established in Proposition 4.10 is needed. Let us consider the following two strategies to draw edges from layer V_j to V_k and from V_k to V_l:

(1) **Strategy 1**: Assign the edges in such a way that the number of incoming edges from V_j and outgoing edges to V_l is as balanced as possible for all nodes of layer V_k; this strategy corresponds to choosing the most balanced splits in $\mathfrak{S}(a^*, n_k, n_j)$ and $\widetilde{\mathfrak{S}}(b^*, n_k, n_l)$.
(2) **Strategy 2**: Assign the edges in such a way that the number of incoming edges from V_j and outgoing edges to V_l is as imbalanced as possible for all nodes of layer V_k; this strategy corresponds to choosing the most imbalanced splits in $\mathfrak{S}(a^*, n_k, n_j)$ and $\widetilde{\mathfrak{S}}(b^*, n_k, n_l)$.

Here we only need Strategy 1, the other strategy will be used further on in Proposition 5.2, where both strategies will be illustrated with some examples. We show that the quantity

$$\frac{1}{n_j n_l} \left\lceil \frac{a^*}{n_k} \right\rceil \left\lceil \frac{b^*}{n_k} \right\rceil$$

acts as an upper bound for the objective function in (15) when Strategy 1 is followed, and a fortiori it will also be an upper bound for the minimum of the integer quadratic program. Strategy 1 corresponds to a way of drawing edges from V_j to V_k and from V_k to V_l such that the third condition in Definition 4.3 is satisfied, because then at most $\lceil \frac{a^*}{n_k} \rceil$ nodes of layer V_j have outgoing edges to V_k and at most $\lceil \frac{b^*}{n_k} \rceil$ nodes have incoming edges from V_k. So, we have at most $\lceil \frac{a^*}{n_k} \rceil \lceil \frac{b^*}{n_k} \rceil$ connected couples through a node of V_k.

Property 6. This property immediately follows from Property 5 since in this case $\frac{a^*}{n_k} = \lceil \frac{a^*}{n_k} \rceil$ and $\frac{b^*}{n_k} = \lceil \frac{b^*}{n_k} \rceil$. □

From these properties it follows that C_{jkl} is a family of discrete conjunctors (the functions are only defined over $\mho_{jk} \times \mho_{kl}$ instead of $[0,1]^2$). It is interesting to look how these functions behave compared to standard t-norms. In Figs. 3–5, we have compared C_{jkl} to respectively $T_{\mathbf{L}}$, $T_{\mathbf{P}}$ and $T_{\mathbf{M}}$ for $n_j = n_k = n_l = 10$. One can see that the function C_{jkl} is always greater than $T_{\mathbf{L}}$ with a peak in the upper triangle. In the region close to $(0,0)$, it is substantially smaller than $T_{\mathbf{P}}$, while it is similar to $T_{\mathbf{P}}$ in the region close to $(1,1)$. Thirdly, in almost all parts of the input domain, C_{jkl} is smaller than $T_{\mathbf{M}}$.

5. ERA ranking representability

Since AUC transitivity acts as a necessary and sufficient condition for AUC ranking representability, it is able to reveal deeper insights of multi-class classifiers, but it is not of great practical value. The functions C_{jkl} are solutions of an integer quadratic program, which is an NP-hard problem [33], and as a result, the condition can only be exactly verified for small data sets. Instead of focussing on intelligent algorithms to solve the integer quadratic program approximately, we will present another approach to circumvent this computational bottleneck. Simultaneously, an analytical expression for the solution of the integer quadratic program is derived.

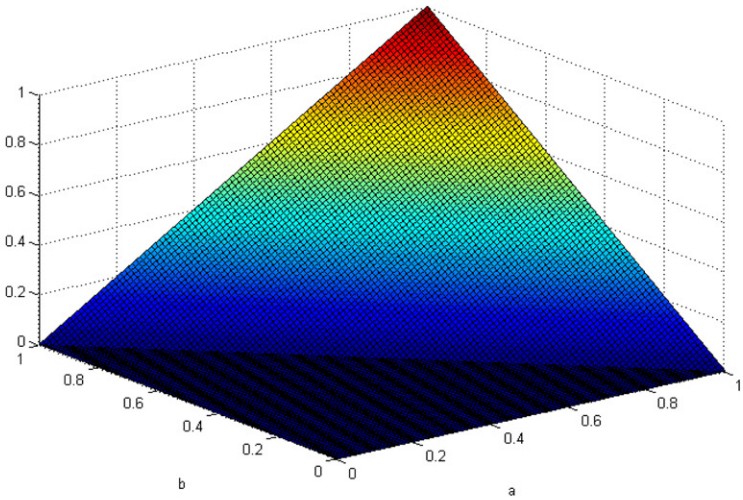
Using the concepts from the previous section, ERA ranking representability naturally follows from AUC ranking representability by considering the abstraction from a finite sample to the underlying distribution. Let us now introduce a specific type of C-transitivity.

Definition 5.1. A reciprocal relation $Q : \mathcal{X}^2 \to [0,1]$ is called ERA-transitive if it is C-transitive w.r.t. the conjunctor $C_{\mathbf{P}0}$ defined by

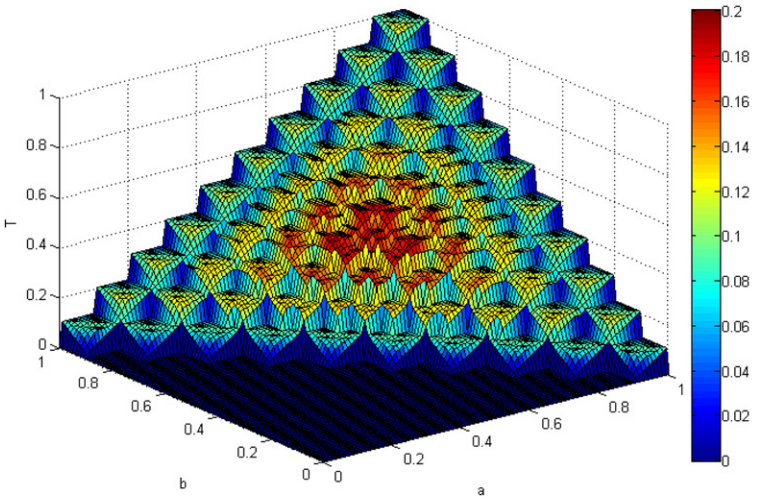
$$C_{\mathbf{P}0}(a,b) = \begin{cases} 0, & \text{if } a+b \leqslant 1, \\ ab, & \text{if } a+b > 1. \end{cases}$$

Remarkably, we can show that ERA transitivity leads to a necessary and sufficient condition for ERA ranking representability.

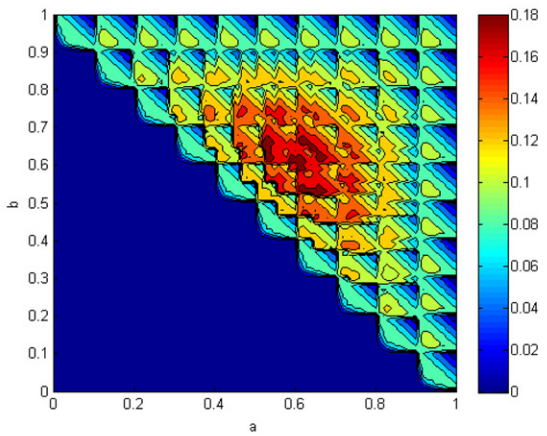
Proposition 5.2. *A triplet $\overline{\mathcal{F}} = \{f_{12}, f_{23}, f_{13}\}$ of bipartite ranking functions is ERA ranking representable on three independent random vectors if and only if the corresponding reciprocal relation of expected ranking accuracies is ERA-transitive.*



(a) $T_{\mathbf{L}}$

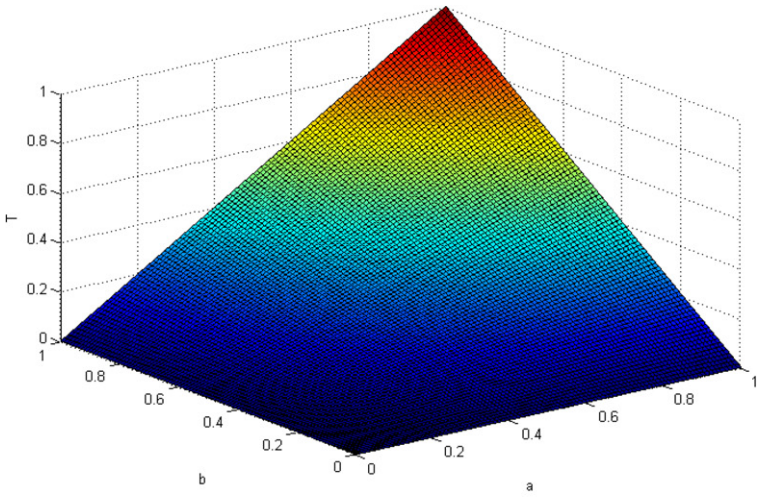


(b) 3D plot of C_{jkl}. The color indicates $C_{jkl} - T_{\mathbf{L}}$.

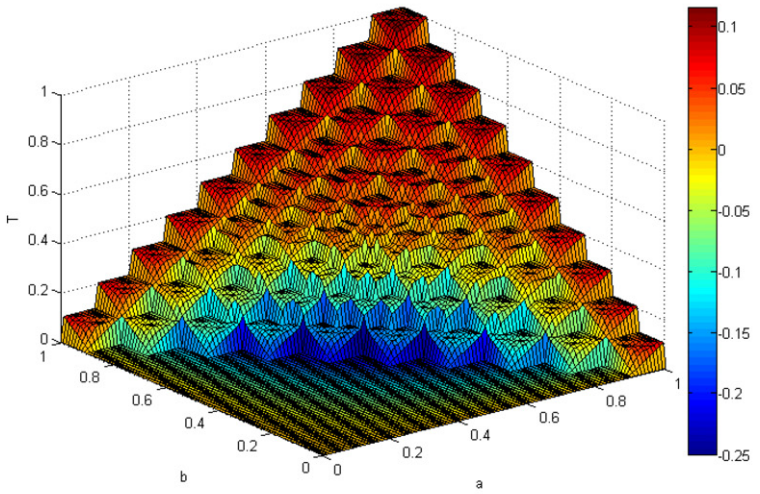


(c) Contour plot of $C_{jkl} - T_{\mathbf{L}}$.

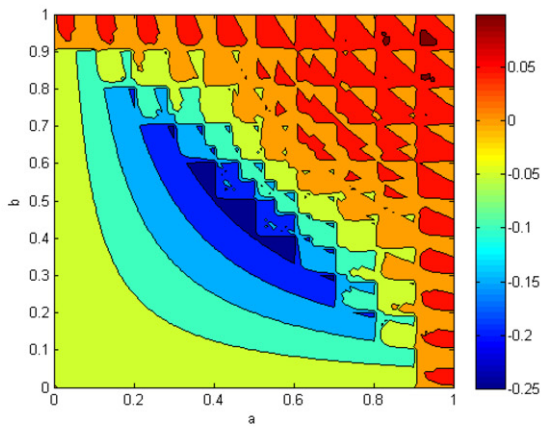
Fig. 3. C_{jkl} compared to $T_{\mathbf{L}}$ with $n_j = n_k = n_l = 10$. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)



(a) $T_{\mathbf{P}}$

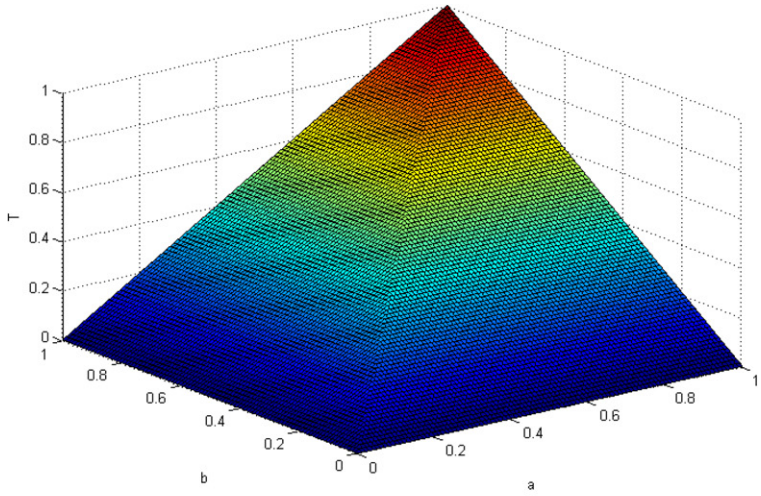


(b) 3D plot of C_{jkl}. The color indicates $C_{jkl} - T_{\mathbf{P}}$.

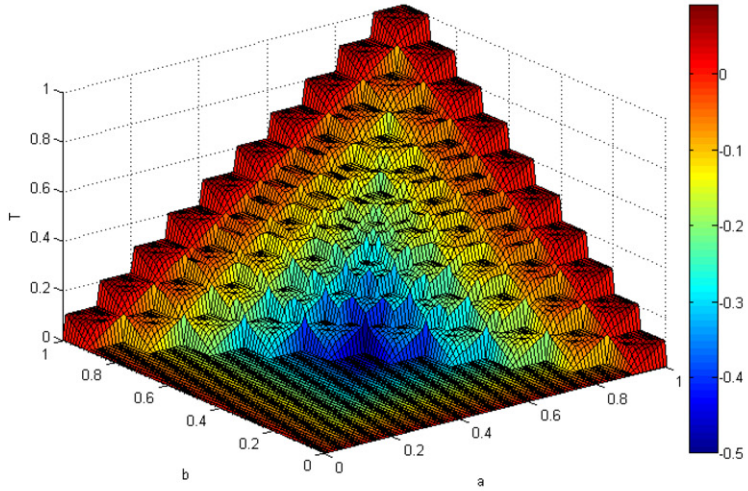


(c) Contour plot of $C_{jkl} - T_{\mathbf{P}}$.

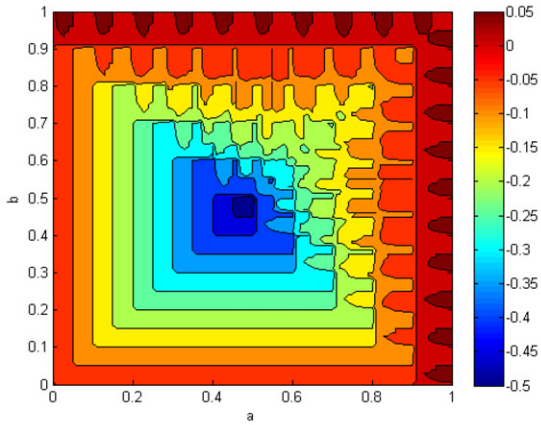
Fig. 4. C_{jkl} compared to $T_{\mathbf{P}}$ with $n_j = n_k = n_l = 10$. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)



(a) $T_{\mathbf{M}}$



(b) 3D plot of C_{jkl}. The color indicates $C_{jkl} - T_{\mathbf{M}}$.



(c) Contour plot of $C_{jkl} - T_{\mathbf{M}}$.

Fig. 5. C_{jkl} compared to $T_{\mathbf{M}}$ with $n_j = n_k = n_l = 10$. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Proof. The proof is inspired by the sufficient condition derived for AUC ranking representability in the previous section. Given that expected ranking accuracy is the immediate generalization of the AUC from a finite sample to the underlying distribution, we only need to show that C_{jkl} converges in the limit to $C_{\mathbf{P}0} : [0,1]^2 \to [0,1]$. In order to examine this limit behavior, we first assume that the ratio of the number of data objects sampled from the respective classes remains unchanged. We will give a formal discussion for the case where $n_j = n_k = n_l$. For other ratios, a more difficult formal proof can be formulated. In the second stage, we construct such a formal proof from the insights gained for the case $n_j = n_k = n_l$.

Case 1. $n_j = n_k = n_l$.

In this case, we have to show that

$$C_{\mathbf{P}0}(a,b) = \lim_{n_j \to \infty} C_{jjj}(a,b),$$

with $n_j \in \mathbb{N}$. We immediately find that in the limit properties (1)–(4) in Proposition 4.15 respectively reduce to monotonicity, commutativity, absorbing element 0 and neutral element 1. However, $C_{\mathbf{P}0}$ will not be a t-norm since associativity does not hold. To compute the limit of C_{jjj}, we will consider three cases.

Subcase 1. $\lfloor \frac{a^*}{n_j} \rfloor + \lfloor \frac{b^*}{n_j} \rfloor > n_j$.

We show that in this case the minimum of the integer quadratic program is found by applying Strategy 1 as described in the proof of the fifth property in Proposition 4.15. The easiest way to recognize this is by considering the graph-theoretic interpretation of Proposition 4.10. Given $\lfloor \frac{a^*}{n_j} \rfloor + \lfloor \frac{b^*}{n_j} \rfloor > n_j$, always paths will be found from V_j to V_l that pass through a node of V_k. The first thing to observe is that the splits for $\mathbf{a}^* = (a_1^*, \ldots, a_{n_j}^*)$ and $\mathbf{b}^* = (b_1^*, \ldots, b_{n_j}^*)$ considered in Strategy 2 lead to a value of 1 for the objective function, because a node in V_k can be found that has incoming edges from all V_j-nodes and outgoing edges to all V_l-nodes. As a consequence, we connect all couples of nodes from $V_j \times V_l$ in this way. Irrespective the split of a^* and b^* that is chosen, we will always find connected couples. The only chance to end up with as few connected couples as possible is by constructing as many paths as possible through couples that have to be connected anyway. This is exactly what is accomplished by Strategy 1, leading to vectors $\mathbf{a}^*$ and $\mathbf{b}^*$ that are constructed as follows:

$$a_i^* = \begin{cases} \lfloor \frac{a^*}{n_j} \rfloor, & \text{if } i \leqslant n_j - a^* \bmod n_j, \\ \lceil \frac{a^*}{n_j} \rceil, & \text{if } i > n_j - a^* \bmod n_j, \end{cases}$$

$$b_i^* = \begin{cases} \lfloor \frac{b^*}{n_j} \rfloor, & \text{if } i > b^* \bmod n_j, \\ \lceil \frac{b^*}{n_j} \rceil, & \text{if } i \leqslant b^* \bmod n_j. \end{cases}$$

Let us now try to derive a closed form for the objective function. By applying Strategy 1, we minimize the number of nodes from V_j with outgoing edges to V_k. At least $\lceil \frac{a^*}{n_j} \rceil$ nodes from V_j have outgoing edges to V_k. $\lfloor \frac{a^*}{n_j} \rfloor$ of these nodes have n_j outgoing edges to V_k (more precisely, to all elements of V_k). Only one node can have less than n_j edges (when a^*/n_j is not an integer). Similarly, we find that at least $\lceil \frac{b^*}{n_j} \rceil$ nodes from V_l have incoming edges from V_k. $\lfloor \frac{b^*}{n_j} \rfloor$ of these nodes have n_j incoming edges, and the remaining node can have less than n_j edges (when b^*/n_j is not an integer). Putting everything together, we find that all V_j-nodes with outgoing edges to V_k and all V_l-nodes with incoming edges from V_k have to be connected in this way, except the V_j-node and V_l-node with less than n_j outgoing (respectively incoming) edges. One can easily verify that these two nodes will also be connected when $a^* \bmod n_j + b^* \bmod n_j \geqslant 1$. This corresponds to the following value for the objective function:

$$\tau_1 = \begin{cases} \frac{1}{n_j^2}(\lceil \frac{a^*}{n_j} \rceil \lceil \frac{b^*}{n_j} \rceil - 1), & \text{if } a^* \bmod n_j + b^* \bmod n_j < 1, \\ \frac{1}{n_j^2} \lceil \frac{a^*}{n_j} \rceil \lceil \frac{b^*}{n_j} \rceil, & \text{if } a^* \bmod n_j + b^* \bmod n_j \geqslant 1. \end{cases} \tag{17}$$

Remark that the -1 corresponds to the couple that is potentially not connected. τ_1 reduces in the limit to the following simple expression

$$\lim_{n_j \to \infty} \tau_1 = \frac{a^* b^*}{n_j^4} = ab.$$

Fig. 6 shows the obtained graph when Strategies 1 and 2 are applied to an example that satisfies Subcase 1.

Subcase 2. $\lceil \frac{a^*}{n_j} \rceil + \lceil \frac{b^*}{n_j} \rceil \leqslant n_j$.

We show that in this case the minimum of the integer quadratic program is 0. This minimum is found by applying Strategy 2, as described in the proof of the fifth property in Proposition 4.15. The vectors $\mathbf{a}^* = (a_1^*, \ldots, a_{n_j}^*)$ and $\mathbf{b}^* = (b_1^*, \ldots, b_{n_j}^*)$ are now constructed as follows:

$$a_i^* = \begin{cases} 0, & \text{if } i < n_j - \lfloor \frac{a^*}{n_j} \rfloor, \\ a^* \bmod n_j, & \text{if } i = n_j - \lfloor \frac{a^*}{n_j} \rfloor, \\ n_j, & \text{if } i > n_j - \lfloor \frac{a^*}{n_j} \rfloor, \end{cases} \tag{18}$$

and

$$b_i^* = \begin{cases} n_j, & \text{if } i < \lceil \frac{b^*}{n_j} \rceil, \\ b^* \bmod n_j, & \text{if } i = \lceil \frac{b^*}{n_j} \rceil, \\ 0, & \text{if } i > \lceil \frac{b^*}{n_j} \rceil, \end{cases} \tag{19}$$

which of course results in a feasible solution for the integer quadratic program. Given that $\lceil \frac{a^*}{n_j} \rceil + \lceil \frac{b^*}{n_j} \rceil \leqslant n_j$, it follows that $a_i^* b_i^* = 0$ and $a_{i-1}^* b_i^* = 0$ for all $i = 1, \ldots, n_j$ such that the objective function becomes zero.

Fig. 7 shows the obtained graph when Strategies 1 and 2 are applied to an example that satisfies Subcase 2.

Subcase 3. $\lfloor \frac{a^*}{n_j} \rfloor + \lfloor \frac{b^*}{n_j} \rfloor \leqslant n_j < \lceil \frac{a^*}{n_j} \rceil + \lceil \frac{b^*}{n_j} \rceil$ (none of the above cases holds).

Besides these two cases, normally also a third case has to be distinguished, when none of the above two conditions holds. However, we do not have to discuss this third case, for which the minimum of the objective function is more difficult to express. Fortunately, this case vanishes in the limit, since

$$\lim_{n_j \to \infty} \left\lceil \frac{a^*}{n_j} \right\rceil - \left\lfloor \frac{a^*}{n_j} \right\rfloor = 0, \qquad \lim_{n_j \to \infty} \left\lceil \frac{b^*}{n_j} \right\rceil - \left\lfloor \frac{b^*}{n_j} \right\rfloor = 0.$$

In Subcase 3, both Strategies 1 and 2 can deliver the minimum of the integer quadratic program, yet it depends on the actual values of a and b whether Strategies 1 or 2 should be applied. Fig. 8 shows the obtained graph when Strategies 1 and 2 are applied on an example that satisfies Subcase 3. For Strategy 1 one can easily see that Eq. (17) still provides the obtained value for the objective function. For Strategy 2, contrary to Subcase 2, the value for the objective function will no longer be zero. Given Subcase 3, one will always find exactly one node in the layer V_k with incoming and outgoing nodes (this means that a_i^* and b_i^* are both different from zero for that node). The values of a_i^* and b_i^* are respectively given by $a^* \bmod n_j$ and $b^* \bmod n_j$. As a consequence, we obtain by applying Eqs. (18) and (19)

$$\tau_2 = \big(a^* \bmod n_j\big)\big(b^* \bmod n_j\big),$$

as value for the objective function in Subcase 3, when Strategy 2 is employed.

General case. $n_j = n_k = n_l$ does not hold.

The proof given above for $n_j = n_k = n_l$ can be easily extended to other cases, while still assuming that the ratio of the number of data objects sampled from the respective classes remains unchanged when the sample size grows to infinity.

Subcase 1. $\lfloor \frac{a^*}{n_j} \rfloor + \lfloor \frac{b^*}{n_l} \rfloor > n_k$.

We have to apply Strategy 1 to obtain the minimum. It is now given by:

$$\tau_1 = \begin{cases} \frac{1}{n_j n_l}(\lceil \frac{a^*}{n_k} \rceil \lceil \frac{b^*}{n_k} \rceil - 1), & \text{if } a^* \bmod n_k + b^* \bmod n_k < 1, \\ \frac{1}{n_j n_l} \lceil \frac{a^*}{n_k} \rceil \lceil \frac{b^*}{n_k} \rceil, & \text{if } a^* \bmod n_k + b^* \bmod n_k \geqslant 1. \end{cases} \tag{20}$$

Subcase 2. $\lceil \frac{a^*}{n_j} \rceil + \lceil \frac{b^*}{n_l} \rceil \leqslant n_k$.

We have to apply Strategy 2 to obtain the minimum. One can easily see that in this case the minimum of the objective function again becomes zero, since the vectors $\mathbf{a}^* = (a_1^*, \ldots, a_{n_j}^*)$ and $\mathbf{b}^* = (b_1^*, \ldots, b_{n_j}^*)$ are now constructed as follows:

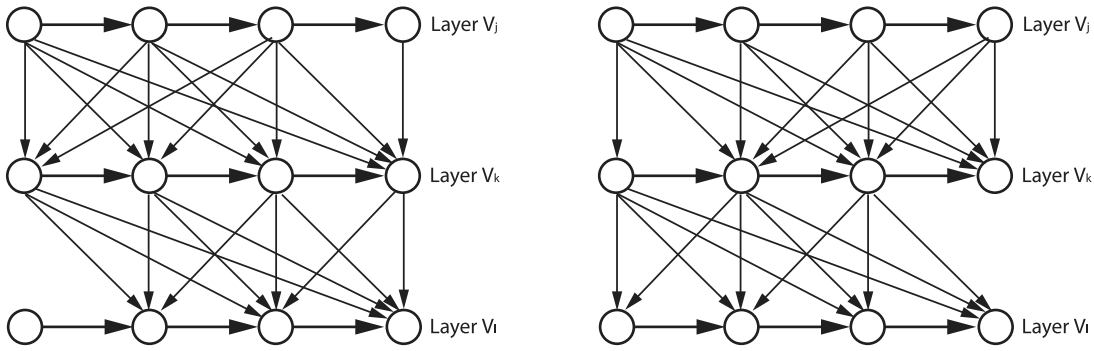


Fig. 6. An example to illustrate Subcase 1 in the proof of Proposition 5.2. Let us consider $n_j = n_k = n_l = 4$, $a = 13/16$ and $b = 11/16$. So, we have to draw 13 edges from layer V_j to layer V_k and 11 edges from layer V_k to layer V_l. The graphs obtained by applying Strategy 1 (left) and Strategy 2 (right) are shown, respectively leading to 11 and 16 connected couples of nodes from layer V_j to layer V_l. One can clearly see that in this example the number of connected couples of nodes from V_j to V_l cannot be lower than 10. We also remark that edges in the opposite direction are left out to simplify the graph, but the graphs are essentially complete graphs, so the 5 edges from V_k to V_j and the 3 edges from V_l to V_k are left out.

$$a_i^* = \begin{cases} 0, & \text{if } i < n_k - \lfloor \frac{a^*}{n_j} \rfloor, \\ a^* \bmod n_j, & \text{if } i = n_k - \lfloor \frac{a^*}{n_j} \rfloor, \\ n_j, & \text{if } i > n_k - \lfloor \frac{a^*}{n_j} \rfloor, \end{cases}$$

and

$$b_i^* = \begin{cases} n_l, & \text{if } i < \lceil \frac{b^*}{n_l} \rceil, \\ b^* \bmod n_l, & \text{if } i = \lceil \frac{b^*}{n_l} \rceil, \\ 0, & \text{if } i > \lceil \frac{b^*}{n_l} \rceil. \end{cases}$$

Subcase 3. $\lfloor \frac{a^*}{n_j} \rfloor + \lfloor \frac{b^*}{n_l} \rfloor \leqslant n_k < \lceil \frac{a^*}{n_j} \rceil + \lceil \frac{b^*}{n_l} \rceil$ (none of the above cases holds).

This case again vanishes in the limit, so that it does not have to be considered further. For the sake of completeness, we also give here the expression for the minimum of the integer quadratic program. As illustrated in Fig. 8, both Strategies 1 and 2 can deliver the minimum of the integer quadratic program. This holds also for the case where $n_j = n_k = n_l$ does not hold. If Strategy 1 is applied, then the objective function takes the value given by Eq. (20). If Strategy 2 is applied, then again exactly one node from the layer V_k will simultaneously have incoming and outgoing edges for Subcase 3. The values of a_i^* and b_i^* are now respectively given by $a^* \bmod n_j$ and $b^* \bmod n_l$. Consequently, the value for the objective function becomes:

$$\tau_2 = \big(a^* \bmod n_j\big)\big(b^* \bmod n_l\big), \tag{21}$$

for Subcase 3, when Strategy 2 is applied. □

The conjunctor $C_{\mathbf{P0}}$ is visualized in Fig. 11(a). It can be expressed as a special type of cycle transitivity, by applying Proposition 3.6.

Proposition 5.3. *A reciprocal relation $Q : \mathcal{X}^2 \to [0,1]$ is ERA-transitive if and only if it is cycle-transitive w.r.t. the upper bound function*

$$U_{C_{\mathbf{P0}}}(\alpha,\beta,\gamma) = \min\big(\alpha + \beta - C_{\mathbf{P0}}(\alpha,\beta), \alpha + \gamma - C_{\mathbf{P0}}(\alpha,\gamma), \beta + \gamma - C_{\mathbf{P0}}(\beta,\gamma)\big).$$

Proposition 5.4. *ERA transitivity implies moderate product transitivity and therefore also dice transitivity.*

Proof. Let us consider a unit square for the couple (β,γ), as visualized in Fig. 9. From the constraint

$$\beta \leqslant \gamma, \tag{22}$$

follows that only the region above the bisector must be considered (i.e., the line given by the identity function). The above upper bound function can be expressed as $\alpha\gamma \leqslant 1 - \beta$. Let us try to make this constraint as tight as possible by choosing $\alpha = \beta$. This means that our cycle transitivity property only imposes a constraint when the following inequality holds:

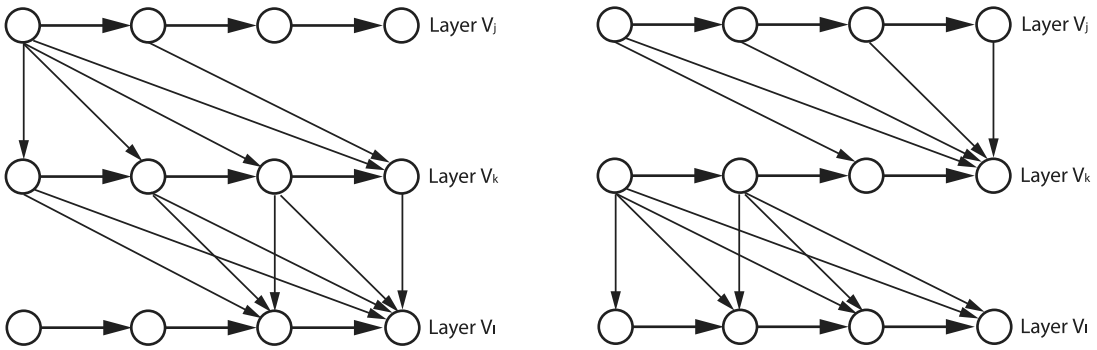


Fig. 7. An example to illustrate Subcase 2 in the proof of Proposition 5.2. Let us consider $n_j = n_k = n_l = 4$, $a = 5/16$ and $b = 7/16$. So, we have to draw 5 edges from layer V_j to layer V_k and 7 edges from layer V_k to layer V_l. The graphs obtained by applying Strategy 1 (left) and Strategy 2 (right) are shown, respectively leading to 3 and 0 connected couples of nodes from layer V_j to layer V_l. We also remark that edges in the opposite direction are left out to simplify the graph, but the graphs are essentially complete graphs, so the 11 edges from V_k to V_j and the 9 edges from V_l to V_k are left out.

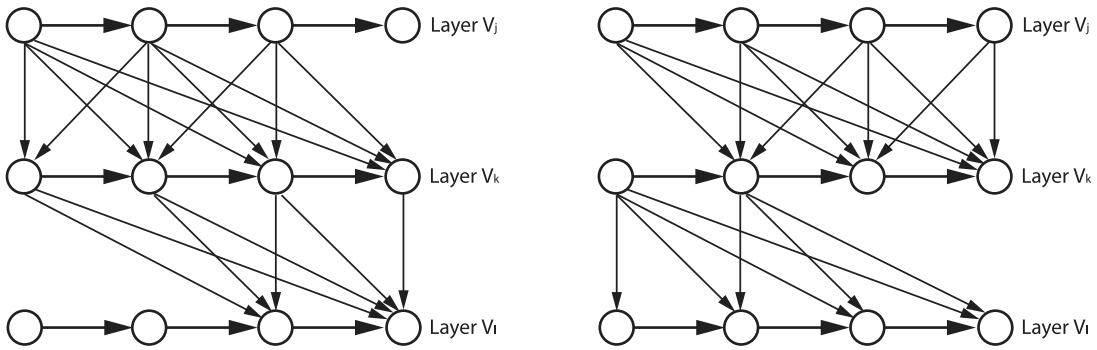


Fig. 8. An example to illustrate Subcase 3 in the proof of Proposition 5.2. Let us consider $n_j = n_k = n_l = 4$, $a = 11/16$ and $b = 7/16$. So, we have to draw 10 edges from layer V_j to layer V_k and 7 edges from layer V_k to layer V_l. The graphs obtained by applying Strategy 1 (left) and Strategy 2 (right) are shown, respectively leading to 5 and 9 connected couples of nodes from layer V_j to layer V_l. We also remark that edges in the opposite direction are left out to simplify the graph, but the graphs are essentially complete graphs, so the 6 edges from V_k to V_j and the 9 edges from V_l to V_k are left out. For this example, Strategy 1 turns out to lead to the minimum of the integer quadratic program, but one can observe that small changes in the values for a and b (such as decreasing b to 5/16) will result in obtaining the minimum with Strategy 2 (3 connected couples versus still 6 with Strategy 1).

$$\gamma > \frac{1-\beta}{\beta}. \tag{23}$$

Moreover, the upper bound function $U(\alpha, \beta, \gamma) = \alpha + \gamma - \alpha\gamma$ neither imposes a constraint when $\alpha + \gamma \leqslant 1$. This corresponds to the region

$$\beta + \gamma \leqslant 1, \tag{24}$$

because we are already assuming that $\alpha = \beta$. Putting everything together, the upper bound function $U(\alpha, \beta, \gamma) = \alpha + \gamma - \alpha\gamma$ only imposes a constraint in the subregion of the unit square defined by inequalities (22), (23) and (24). We have visualized this region in Fig. 9(a) with a gray background color. Obviously, $C_{\mathbf{P0}}(\beta, \gamma)$ equals $\beta\gamma$ in this part of the unit square. ERA transitivity is therefore a stronger type of transitivity than cycle transitivity w.r.t. the upper bound function $U(\alpha, \beta, \gamma) = \alpha + \gamma - \alpha\gamma$. □

This proposition mainly confirms that all pieces of the puzzle fit surprisingly well. In the previous sections it was shown how AUC transitivity induces a sufficient condition for AUC ranking representability, while dice transitivity could only lead to a necessary condition. From this we were able to prove indirectly that the former type of transitivity had to be stronger than the latter one, but this could not be observed directly from the upper bound functions. Since this relationship between both types of cycle transitivity can be observed very easily in the infinite case, it gives an additional confirmation of the correctness of our analysis in the finite case. To draw the attention of the reader to the potentially tight bound between AUC transitivity and dice transitivity, Fig. 10 visualizes all considered regions for $n_j = n_k = n_l = 10$. It is shown that for this choice of n_j, n_k, n_l the region defined by inequalities (22)–(24) does not overlap with the region where $C_{jkl}(\beta, \gamma)$ exceeds $\beta\gamma$, albeit both regions are located very close to each other.

Proposition 5.5. *$T_{\mathbf{P}}$-transitivity implies $C_{\mathbf{P0}}$-transitivity.*

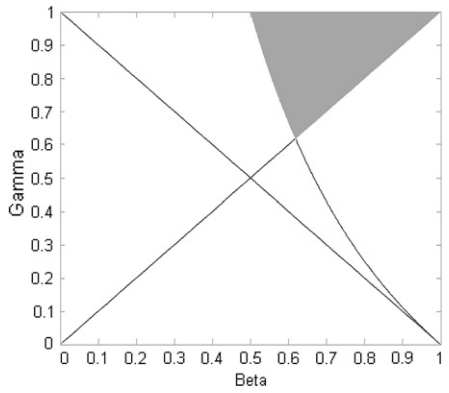


Fig. 9. A visualization of the unit square for (β, γ). The shaded region indicates the zone of the unit square where all three inequalities (22)–(24) hold.

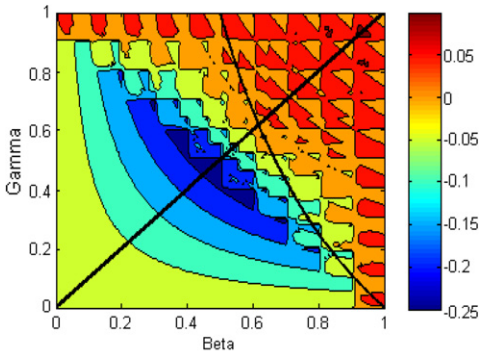


Fig. 10. Contour plot of the difference between AUC transitivity and dice transitivity (or cycle transitivity w.r.t. the upper bound function $U(\alpha, \beta, \gamma) = \alpha + \gamma - \alpha\gamma$), because in the proof the assumption is made that $\alpha = \beta$). One can see that AUC transitivity dominates the other two types of transitivity in the region where all three inequalities (22)–(24) hold (shown for $n_j = n_k = n_l = 10$).

Proof. Immediate since $C_{\mathbf{P}0} \leqslant T_{\mathbf{P}}$. □

6. Practical considerations

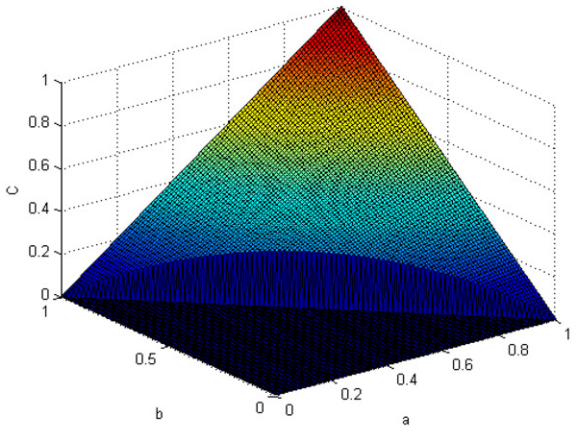
6.1. Verifying ERA or AUC ranking representability

In the previous sections an interesting sufficient condition was established for reducing one-versus-one ensembles to ranking models by investigating the pairwise AUCs. The sufficient condition for the three-class case can be verified by solving an integer quadratic program. This class of problems can in general not be solved exactly in polynomial time. However, we have not elaborated on this issue, since we were able to derive a more simple expression for the infinite case where a generalization is made from a sample to the underlying distribution. As a consequence, it makes sense to verify ERA transitivity on the pairwise AUCs instead of AUC transitivity, since we are mainly interested in generalizing to out-of-sample data. The conjunctor $C_{\mathbf{P}0}$ is visualized in Fig. 11 and compared to C_{jkl} for $n_j = n_k = n_l = 10$. Although such a sample size can be considered as unrealistically small, it turns out that even then $C_{\mathbf{P}0}$ behaves very similarly to C_{jkl}. Thus, the approximation makes sense. Moreover, in the proof of the previous proposition, we have derived an analytical expression for the solution of the integer quadratic program, so that no optimization algorithm is required. This can be summarized as follows.

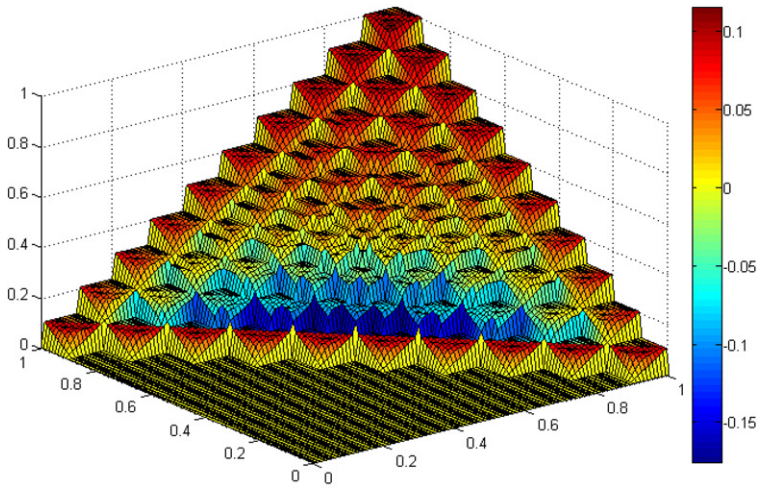
Corollary 6.1. *For any values of n_j, n_k, n_l, a^* and b^*, the solution of integer quadratic program* (15) *can be expressed as*

$$\tau = \begin{cases} \tau_1, & \text{if } \lfloor \frac{a^*}{n_j} \rfloor + \lfloor \frac{b^*}{n_l} \rfloor > n_k, \\ \min(\tau_1, \tau_2), & \text{if } \lfloor \frac{a^*}{n_j} \rfloor + \lfloor \frac{b^*}{n_l} \rfloor \leqslant n_k \leqslant \lceil \frac{a^*}{n_j} \rceil + \lceil \frac{b^*}{n_l} \rceil, \\ 0, & \text{if } \lceil \frac{a^*}{n_j} \rceil + \lceil \frac{b^*}{n_l} \rceil \leqslant n_k, \end{cases}$$

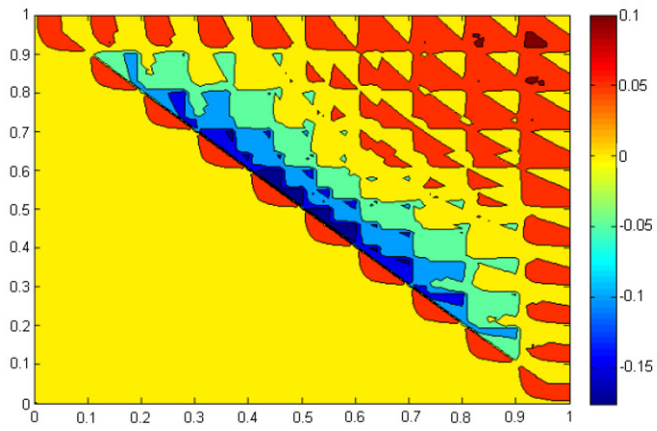
with τ_1 and τ_2 respectively defined by Eqs. (20) *and* (21).



(a) The conjunctor $C_{\mathbf{P0}}$



(b) 3D plot of C_{jkl}. The color indicates $C_{jkl} - C_{\mathbf{P0}}$.



(c) Contour plot of $C_{jkl} - C_{\mathbf{P0}}$.

Fig. 11. C_{jkl} compared to $C_{\mathbf{P0}}$ with $n_j = n_k = n_l = 10$. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

So, we have obtained a verifiable condition to check whether a one-versus-one ensemble can be reduced to a ranking model for the three-class case, but will this approach also work in practice? The answer is not unhesitatingly yes for the following two main reasons:

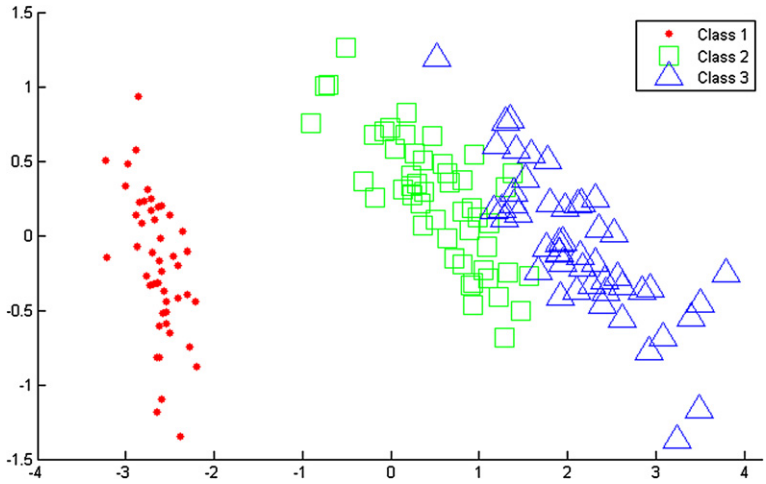


Fig. 12. A plot of the first two principal components of the iris data set.

Table 1
A comparison of the performance obtained on the iris data set by ten-fold cross-validation with various multi-class methods and a simple ranking model. The results of the multi-class methods are duplicated from [32] (OVO = one-versus-one SVM, DAG = directed acyclic graphs, OVA = one-versus-all SVM, W&W = single machine approach of Weston and Watkins, C&S = single machine approach of Crammer and Singer, SVORIM = support vector ordinal regression).

Method	OVO	DAG	OVA	W&W	C&S	SVORIM
Acc.	97.333	97.333	96.000	97.333	87.333	**98.000**

(1) The framework of AUC ranking representability only identifies the existence of a single ranking function that yields the same error in terms of pairwise expected ranking accuracy as a one-versus-one ensemble, but nothing can be said about the complexity of this ranking model. When we minimize a loss function over a hypothesis space of ranking models, no guarantee can be given that the model that we try to find is included in this hypothesis space.
(2) Ordinal regression models do not try to optimize accuracy, but a loss function that takes the magnitude of an error into account. However, the magnitude of an error has no meaning at all in a multi-class setting, even if we artificially try to impose an order on the classes. The optimization of a magnitude-based loss function can harm the performance significantly when only accuracy is taken into account as performance measure. This effect can even be observed for ordinal regression data sets.

These two findings imply that in practice an AUC ranking representable one-versus-one model will not always be beaten by a ranking model. Revealing the situations where a performance gain will be obtained might not be done easily and would require an experimental validation on numerous data sets. Since most benchmark problems for multi-class classification consider more than three classes, first an extension of our approach to more than three classes has to be elaborated. An obvious generalization could be established by looking at all triplets of classes and simplifying those for which AUC ranking representability is fulfilled, but further research is required to verify whether this idea would work. Hereunder we have analyzed two three-class benchmark problems from the UCI repository to illustrate the potential benefits of AUC ranking representability. Both problems have been analyzed by [32] in an experimental comparison of different multi-class schemes with SVMs as base classifiers. We decided to compare with their results because they describe their experimental setup in such a way that the experiments could be easily replicated.

6.2. Iris data

The first data set that was analyzed is the well-known *iris* data set, which is probably one of the most frequently utilized data sets to evaluate multi-class classifiers. The first two principal components of the data are visualized in Fig. 12. One can see that class $\mathcal{C}_2$ is sandwiched on the left side by class $\mathcal{C}_1$ and class $\mathcal{C}_3$ on the right side, thus theoretically we impose the order $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \mathcal{C}_3$ on the classes if we would fit an ordinal regression model to that data set. In a comparison paper of kernel-based multi-class classification methods, Hsu and Lin [32] report for this data set that a one-versus-one model outperforms all other multi-class schemes. Since the iris data does not have an accompanying test set, they draw their conclusions based on the 10-fold cross-validation error obtained for the best C (cost parameter) and γ (width of RBF-kernel) found during model selection. A short overview of the results is given in Table 1 together with the results obtained by fitting the kernel-based ordinal regression model of [6] to the data. This method constructs a number of parallel hyperplanes in a high-dimensional space for ordinal regression, similar to the SVM for binary classification. In an initial stage, we first fitted

Table 2
A comparison of the performance obtained on the DNA data set (independent test set) with various multi-class methods and a simple ranking model. The results of [32] are given in the first row and our results in the second row.

Method	OVO	DAG	OVA	W&W	C&S	SVORIM
Acc.	95.441	95.447	95.784	95.618	95.889	
Acc.	94.266					76.560

a one-versus-one SVM to the data by using the values of the hyperparameters specified by [32], resulting in the following pairwise AUCs when the whole data set is used for training[2]:

$$\widehat{A}_{12} = 1.0, \qquad \widehat{A}_{23} = 0.995, \qquad \widehat{A}_{13} = 1.0.$$

These pairwise AUCs definitely satisfy $C_{\mathbf{P}_0}$-transitivity, so a reduction to a ranking model would make sense theoretically. When we fit an ordinal regression model to the data with the SVORIM-package, then the following pairwise AUCs are measured on training data:

$$\widehat{A}_{12} = 1.0, \qquad \widehat{A}_{23} = 0.998, \qquad \widehat{A}_{13} = 1.0.$$

Thus, the performance on training data increases, but more importantly, a better cross-validated performance in terms of accuracy is obtained with the ordinal regression model, using the same experimental setup as [32]. This might be surprising at first sight, but one must take into account that PCA analysis already identified an ordinal structure of the classes. This is clearly an example where a reduction to a single ranking model can improve the generalization performance.

6.3. DNA data

The second data set that was analyzed is the DNA data set. Contrary to the iris data set, this data set has substantially more instances. The data is also relatively high-dimensional (180 features) such that the curse of dimensionality can play a role. Furthermore, the data has been split into a train and test set, so one can avoid the use of cross-validation here. Using the methodology of [32], this resulted in the following pairwise AUCs on the training set:

$$\widehat{A}_{12} = 1.0, \qquad \widehat{A}_{23} = 0.952, \qquad \widehat{A}_{13} = 0.909.$$

This triplet of pairwise AUCs again results in an AUC ranking representable model and suggests the order $\mathcal{C}_1 < \mathcal{C}_3 < \mathcal{C}_2$ on the classes. So, let us swap classes $\mathcal{C}_2$ and $\mathcal{C}_3$, then the following pairwise AUCs on training data are obtained:

$$\widehat{A}_{12} = 0.909, \qquad \widehat{A}_{23} = 0.952, \qquad \widehat{A}_{13} = 1.0.$$

Subsequently, we tested the one-versus-one SVM and the SVORIM algorithm on the test set. Using the same methodology as [32], we were not able to achieve exactly the same results for this data set, yet we obtained a similar (but slightly worse) accuracy on the test set. We cannot give any plausible explanation since we adopted exactly the same setup. The results are summarized in Table 2 and give the impression that the SVORIM algorithm is not able to compete with the other multi-class approaches. However, nothing is further from the truth, as the opposite conclusion can be drawn from the pairwise AUCs measured on the test set. The following values are obtained for the one-versus-one model:

$$\widehat{A}_{12} = 0.808, \qquad \widehat{A}_{23} = 0.833, \qquad \widehat{A}_{13} = 0.997,$$

while the ordinal regression model yields substantially better pairwise AUCs on the test set:

$$\widehat{A}_{12} = 0.906, \qquad \widehat{A}_{23} = 0.907, \qquad \widehat{A}_{13} = 0.996.$$

So, how can these surprisingly different trends between accuracy on the one hand and the pairwise AUCs on the other hand be explained? As discussed above, this is caused by the fact the SVORIM algorithm does not optimize accuracy but a magnitude-based loss function. Fig. 13 gives a good overview of what is going on. On the left side, it shows the first two principal components of the test set with the real labels and on the right side it shows the same test set with the labels predicted by SVORIM. The ordinal regression algorithm clearly assigns too many instances to the middle class in an attempt to minimize the magnitude of errors. Apparently, that does not affect the performance in terms of pairwise AUCs for the DNA data set. It is therefore definitely recommended to look at this last criterion instead of accuracy in order to compare the performance of one-versus-one models and single ranking models.

[2] Remark that we have to train on the whole data set in order to compute pairwise AUCs, since multivariate performance measures cannot be computed unambiguously by means of cross-validation [3].

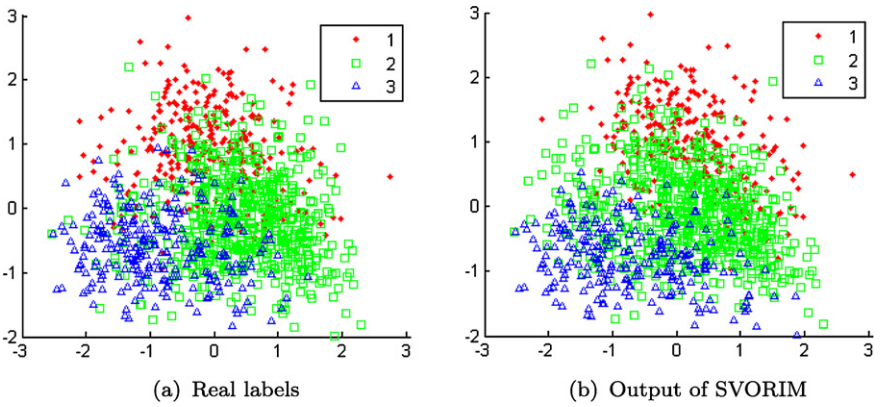


Fig. 13. A graphical illustration of the effect of minimizing a magnitude-based loss function with the SVORIM algorithm on the DNA test set. The first two principal components are shown with the real labels on the left side and the predicted labels on the right side. As a result of the transitivity analysis of the pairwise AUCs, the labels of classes $\mathcal{C}_2$ and $\mathcal{C}_3$ were swapped.

7. Conclusion

In this article we analyzed the transitivity of expected ranking accuracy, a reciprocal preference relation that can be constructed in a pairwise multi-class setting. Similar to the utility representability of more common reciprocal preference relations, this relation naturally led to the concept of ERA ranking representability for bipartite ranking functions constructed on couples of random vectors. In order to find necessary and sufficient conditions for this type of representability, we first recapitulated results obtained for the finite sample case, for which expected ranking accuracy can be estimated by the AUC. ERA ranking representability then reduces to AUC ranking representability, for which necessary and sufficient conditions could be found, based on a graph-theoretic reformulation of the problem and a new type of transitivity, namely AUC transitivity. In previous work we showed that this new type of transitivity can be verified by solving an integer quadratic program. In this article we proved some interesting properties of AUC transitivity, and we generalized AUC ranking representability to ERA ranking representability by analyzing the limit behavior of AUC transitivity. In this way, a distribution-independent and easily verifiable condition was obtained for the three-class case. Extensions for more than three classes are currently under development but invoke a strongly increasing complexity to the problem, as cycle transitivity is only defined on three-element sets at this moment. A generalization of cycle transitivity is therefore the first candidate for future work. Some initial research confirmed that such a generalization is feasible.

From a machine learning point of view, we investigated whether a pairwise multi-class classification model can be simplified to a ranking model (an ordinal regression model to be more precise). To this end, we started from the assumption that the optimal complexity of a multi-class classifier is problem-specific (data-dependent). Reducing a pairwise multi-class classifier to an ordinal regression model can be seen as a quite drastic application of the bias-variance trade-off: a pairwise multi-class classifier is complex, containing many parameters that result in a low bias and a high variance of the performance, while an ordinal regression model contains substantially less parameters, leading to a high bias, but a low variance. So, we did not claim that a pairwise multi-class classifier can always be reduced to an ordinal regression model, we rather looked for necessary and sufficient conditions that allow for such a reduction, by analyzing the pairwise expected ranking accuracies. The result that we obtained is in this regard remarkable and important, as it confirms that the optimal complexity of a multi-class classification model depends on the distribution of the data. The conditions that we derived are moreover distribution-independent, which means that they hold for any distribution of the data.

ERA ranking representability cannot be verified since the distribution of the data is usually unknown. Nevertheless, by evaluating $C_{\mathbf{P}0}$-transitivity on the pairwise AUCs, we have obtained a verifiable condition to check whether a one-versus-one ensemble can be reduced to an ordinal regression model for the three-class case. This relaxation makes sense, because the main interest is a good generalization performance. Simultaneously, we also derived a closed-form expression for the solution of the integer quadratic program, so that both an approximation or a time consuming combinatorial optimization can be avoided.

In practice an ERA ranking representable set of bipartite ranking functions will not always be beaten by a single ranking function. To this end, new machine learning algorithms have to be developed. Initial experimental results on two toy problems illustrate that the reduction to a single ranking model can improve the performance of an algorithm, definitely in terms of pairwise AUCs, but not necessarily in terms of accuracy. However, this is only the start of further experimental research on this topic, in which the obvious question of generalizing ranking representability to more than three classes needs to be tackled. We refer to future work for this extension and an in-depth empirical validation.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. Willem Waegeman would like to thank Eyke Hüllermeier, Hendrik Blockeel and Hans De Meyer for a fruitful discussion about this work on the occasion of his PhD defence. Willem Waegeman is currently supported by the Research Foundation – Flanders. For a part of this work, he has been previously supported by a grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

References

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, D. Roth, Generalization bounds for the area under the ROC curve, Journal of Machine Learning Research 6 (2005) 393–425.
[2] A. Agresti, Categorical Data Analysis, 2nd version, John Wiley and Sons, 2002.
[3] A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, T. Salakoski, An experimental comparison of cross-validation techniques for estimating the area under the ROC curve, Computational Statistics and Data Analysis, 2009, submitted for publication.
[4] A. Billot, An existence theorem for fuzzy utility functions: A new elementary proof, Fuzzy Sets and Systems 74 (1995) 271–276.
[5] U. Bodenhofer, B. De Baets, J. Fodor, A compendium of fuzzy weak orders, Fuzzy Sets and Systems 158 (2007) 811–829.
[6] W. Chu, S. Keerthi, Support vector ordinal regression, Neural Computation 19 (3) (2007) 792–815.
[7] S. Clémençon, N. Vayatis, Ranking the best instances, Journal of Machine Learning Research 8 (2007) 2671–2699.
[8] C. Cortes, M. Mohri, AUC optimization versus error rate minimization, in: Advances in Neural Information Processing Systems 16, Vancouver, Canada, 2003, MIT Press, 2003, pp. 313–320.
[9] K. Crammer, Y. Singer, Pranking with ranking, in: Proceedings of the Conference on Neural Information Processing Systems, Vancouver, Canada, 2001, pp. 641–647.
[10] B. De Baets, H. De Meyer, Transitivity frameworks for reciprocal relations: Cycle-transitivity versus *FG*-transitivity, Fuzzy Sets and Systems 152 (2005) 249–270.
[11] B. De Baets, H. De Meyer, B. De Schuymer, S. Jenei, Cyclic evaluation of transitivity of reciprocal relations, Social Choice and Welfare 26 (2006) 217–238.
[12] B. De Baets, B. De Schuymer, H. De Meyer, Cycle-transitive comparison of artificially coupled random variables, International Journal of Approximate Reasoning 47 (2008) 306–322.
[13] B. De Baets, H. De Meyer, K. De Loof, On the cycle transitivity of the mutual rank probability relation of a poset, Fuzzy Sets and Systems 161 (2010) 2695–2708, doi:10.1016/j.fss.2010.05.005.
[14] H. De Meyer, B. De Baets, B. De Schuymer, On the transitivity of the comonotonic and countermonotonic comparison of random variables, Journal of Multivariate Analysis 98 (2007) 177–193.
[15] B. De Schuymer, H. De Meyer, B. De Baets, S. Jenei, On the cycle-transitivity of the dice model, Theory and Decision 54 (2003) 261–285.
[16] B. De Schuymer, H. De Meyer, B. De Baets, Cycle-transitive comparison of independent random variables, Journal of Multivariate Analysis 96 (2005) 352–373.
[17] B. De Schuymer, H. De Meyer, B. De Baets, Extreme copulas and the comparison of ordered lists, Theory and Decision 62 (2007) 195–212.
[18] J.-P. Doignon, B. Monjardet, M. Roubens, Ph. Vincke, Biorder families, valued relations and preference modelling, Journal of Mathematical Psychology 30 (1986) 435–480.
[19] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.
[20] C. Ferri, J. Hernandez-Orallo, M.A. Salido, Volume under ROC surface for multi-class problems, in: Proceedings of the European Conference on Machine Learning, Dubrovnik, Croatia, 2003, pp. 108–120.
[21] J. Fieldsend, M. Everson, Formulation and comparison of multi-class ROC surfaces, in: Proceedings of the ICML Workshop on ROC Analysis in Machine Learning, Bonn, Germany, 2005, pp. 49–56.
[22] P. Fishburn, Utility Theory for Decision Making, Wiley, 1970.
[23] P. Flach, The geometry of ROC space: Understanding machine learning metrics through ROC isometrics, in: Proceedings of the International Conference on Machine Learning, Washington, DC, USA, 2003.
[24] P. Flach, The many faces of ROC analysis in machine learning, Tutorial Presented at the European Conference on Machine Learning, Valencia, Spain, August 2004.
[25] L. Fono, N. Andjiga, Utility function of fuzzy preferences on a countable set under max-$*$-transitivity, Social Choice and Welfare 28 (2007) 667–683.
[26] J. Fürnkranz, Round robin classification, Journal of Machine Learning Research 2 (2002) 723–747.
[27] J. Fürnkranz, E. Hüllermeier, S. Vanderlooy, Binary decomposition methods for multipartite ranking, Lecture Notes in Computer Science 5781 (2009) 359–374.
[28] D. Hand, R. Till, A simple generalization of the area under the ROC curve for multiple class problems, Machine Learning 45 (2001) 171–186.
[29] J. Hanley, B. McNeil, The meaning and use of the area under a receiver operating characteristics curve, Radiology 143 (1982) 29–36.
[30] T. Hastie, R. Tibshirani, Classification by pairwise coupling, The Annals of Statistics 26 (2) (1998) 451–471.
[31] R. Herbrich, T. Graepel, K. Obermayer, Large margin rank boundaries for ordinal regression, in: A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, 2000, pp. 115–132.
[32] C. Hsu, C. Lin, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks 13 (2002) 415–425.
[33] Z. Hua, B. Zhang, X. Xu, A new variable reduction technique for convex integer quadratic programs, Applied Mathematical Modelling 32 (2008) 224–231.
[34] E. Hüllermeier, J. Fürnkranz, Pairwise preference learning and ranking, in: Proceedings of the European Conference on Machine Learning, Dubrovnik, Croatia, 2003, pp. 145–156.
[35] E. Hüllermeier, J. Hühn, Is an ordinal class structure useful in classifier learning? International Journal of Data Mining, Modelling and Management 1 (1) (2009) 45–67.
[36] M. Koppen, Random utility representation of binary choice probabilities: Critical graphs yielding critical necessary conditions, Journal of Mathematical Psychology 39 (1995) 21–39.
[37] R. Luce, P. Suppes, Handbook of Mathematical Psychology, Preference, Utility and Subjective Probability, Wiley, 1965, pp. 249–410.
[38] P. McCullagh, Regression models for ordinal data, Journal of the Royal Statistical Society, Series B 42 (2) (1980) 109–142.
[39] M. Öztürk, A. Tsoukiàs, Ph. Vincke, Preference modelling, in: J. Figueira, S. Greco, M. Ehrgott (Eds.), Multiple Criteria Decision Analysis. State of the Art Surveys, Springer-Verlag, 2005, pp. 27–71.
[40] F. Provost, T. Fawcett, Robust classification for imprecise environments, Machine Learning 42 (2001) 203–231.
[41] R. Rifkin, A. Klautau, In defense of one-versus-all classification, Journal of Machine Learning Research 5 (2004) 101–143.

[42] A. Shashua, A. Levin, Ranking with large margin principle: Two approaches, in: Advances in Neural Information Processing Systems, vol. 16, Vancouver, Canada, 2003, MIT Press, 2003, pp. 937–944.
[43] Z. Switalski, General transitivity conditions for fuzzy reciprocal preference matrices, Fuzzy Sets and Systems 137 (2003) 85–100.
[44] V. Torra, J. Domingo-Ferrer, J.M. Mateo-Sanz, M. Ng, Regression for ordinal variables without underlying continuous variables, Information Sciences 176 (2006) 465–476.
[45] A. Tversky, Preference, Belief and Similarity, MIT Press, 1998.
[46] W. Waegeman, B. De Baets, A transitivity analysis of bipartite rankings in pairwise multi-class classification, Information Sciences 180 (2010) 4099–4117.
[47] W. Waegeman, B. De Baets, L. Boullart, ROC analysis in ordinal regression learning, Pattern Recognition Letters 29 (2008) 1–9.
[48] F. Wu, C. Lin, R. Weng, Probability estimates for multi-class support vector machines by pairwise coupling, Journal of Machine Learning Research 5 (2004) 975–1005.
[49] L. Yan, R. Dodier, M. Mozer, R. Wolniewiecz, Optimizing classifier performance via an approximation to the Wilcoxon–Mann–Whitney statistic, in: Proceedings of the International Conference on Machine Learning, Washington DC, USA, 2003, pp. 848–855.